

Safety and Risk Pathways in Cooperative Generative Multi-Agent Systems: A Telecom Perspective

Zeinab Nezami, Shehr Bano, Abdelaziz Salama, Maryam Hafeez, Syed Ali Raza Zaidi
School of Electrical and Electronic Engineering, University of Leeds
Leeds, United Kingdom

Abstract

Generative multi-agent systems are rapidly emerging as transformative tools for scalable automation and adaptive decision-making in telecommunications. Despite their promise, these systems introduce novel risks that remain underexplored, particularly when agents operate asynchronously across layered architectures. This paper investigates key safety pathways in telecom-focused Generative Multi-Agent Systems (GMAS), emphasizing risks of miscoordination and semantic drift shaped by persona diversity. We propose a modular safety evaluation framework that integrates agent-level checks on code quality and compliance with system-level safety metrics. Using controlled simulations across 32 persona sets, five questions, and multiple iterative runs, we demonstrate progressive improvements in analyzer penalties and Allocator–Coder consistency, alongside persistent vulnerabilities such as policy drift and variability under specific persona combinations. Our findings provide the first domain-grounded evidence that persona design, coding style, and planning orientation directly influence the stability and safety of telecom GMAS, highlighting both promising mitigation strategies and open risks for future deployment.

CCS Concepts

• **Computing methodologies** → **Multi-agent planning; Simulation evaluation**; • **Networks** → **Wireless access networks**.

Keywords

Generative Multi-Agent Systems, GMAS, Telecommunications, ORAN, Safety Evaluation, Miscoordination

1 Introduction

GMAS offer transformative potential for critical infrastructure by enabling scalable automation, adaptive decision-making, and collaborative intelligence [1]. In telecommunications, these systems promise self-evolving networks [2–4], where agents autonomously perceive, reason, and adjust network behavior in real time. However, GMAS introduce novel classes of risk that differ from those in single-agent or traditional automation frameworks [5], which remain largely unexplored. The scarcity of real-world deployments combined with unresolved safety challenges in even single-agent AI, highlights the urgent need to identify and mitigate emergent risks prior to widespread adoption [5].

As wireless communication evolves toward self-evolving 6G networks, AI-driven systems will autonomously integrate telemetry, user intent, and environmental signals to refine internal policies, decision logic, and workflow across layers [2, 3, 6]. Open Radio Access Network (O-RAN) [7], together with large language model (LLM)-enhanced intelligence [8, 9], enables dynamic, programmable control, allowing networks to adapt to both operational and societal demands. Yet existing multi-agent evaluation frameworks [10–12] largely emphasize social interaction, persona adherence, or game-theoretic reasoning, while overlooking operational risks in high-stakes domains.

Despite their promise, GMAS are vulnerable to three primary failure modes [5]: miscoordination, conflict, and collusion. This paper focuses on miscoordination, which arises when agents act on incomplete or inconsistent information, leading to incompatible actions. For example, in an O-RAN deployment, one agent may reserve spectrum based on local telemetry while another follows a conflicting plan derived from a different state, causing resource contention or dropped connections. Such failures emerge naturally from asynchronous actions, divergent tool outputs, or mismatched knowledge contexts, without requiring adversarial intent. Information asymmetry is particularly pronounced in retrieval-augmented generation (RAG) setups, where agents query distinct document stores or maintain non-overlapping memory contexts. Given their layered architecture, real-time constraints, and safety-critical operations, telecom systems provide a compelling testbed for studying these dynamics.

This work proposes a GMAS for network orchestration and automation in O-RAN, and uses it as a testbed to develop and validate a safety evaluation framework for cooperative GMAS. Our framework evaluates agents that share a common objective while executing modular tasks through distributed roles with asymmetric context. Our contributions are as follows: (i) Safety evaluation framework: We design a domain-agnostic framework that combines agent-level checks with system-level safety metrics, and demonstrate its application in O-RAN and RAN Intelligent Controller (RIC) contexts. (ii) Knowledge–persona integration: We combine heterogeneous knowledge layers (RAG and graph-RAG) with systematic persona variation, and show how these dimensions jointly influence safety, alignment, and reproducibility in multi-agent systems. (iii) Proposed GMAS for telecom orchestration: We implement a cooperative GMAS for O-RAN automation, providing a concrete instantiation of our evaluation framework. (iv) Empirical insights: Through multi-run experiments, we uncover progressive improvements in GMAS performance, agent consistency, and embedding stability. (v) Open resources: We release a GitHub repository containing datasets and code to support reproducibility.



This work is licensed under a Creative Commons Attribution 4.0 International License.
OpenRan '25, November 4-8, 2025, Hong Kong, China
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1977-6/25/11
<https://doi.org/10.1145/3737900.3770174>

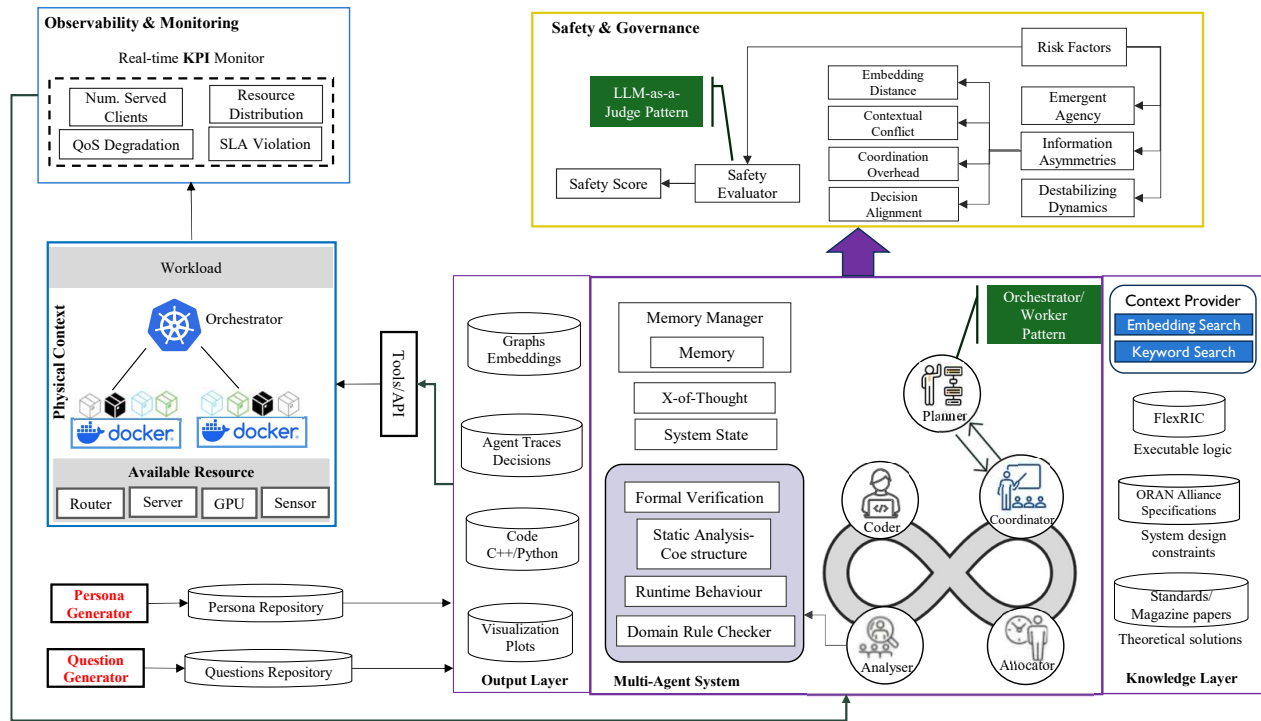


Figure 1: Architecture of safety framework for generative-based multi-agent systems in O-RAN setup

The paper is organised as follows: Section 2 reviews related work, Section 3 presents the GMAS and safety framework, Section 4 reports experiments and results, and Section 5 concludes with future directions.

2 Related Work

Generative AI (GenAI) has demonstrated strong potential in automating telecom and reducing human effort in tasks such as troubleshooting recommendations [13], code refactoring [14], and network configuration [15, 16]. While recent work has explored GMAS for telecom, the emphasis has largely been on benchmarking [17] and automation [18] rather than safety. To the best of our knowledge, there is no dedicated research addressing safety risks in GMAS for telecom. Related discussions have instead focused either on the safety of GenAI in other domains, or on artificial intelligence safety for telecom more generally. We review three strands of related work: (i) GMAS for telecom, (ii) safety in GenAI-based single-agent systems across domains, and (iii) safety evaluation for GMAS frameworks in telecom.

Panek et al. [19] propose a modular framework for cloud-native 5G networks that employs chain-of-thought reasoning before delegating tasks to cloud management tools. Zou et al. [20] introduce an edge-based multi-model approach that deploys LLM on wireless devices rather than a central server, reducing reliance on centralized infrastructure but increasing computational overhead. Similarly, Li et al. [21] design a multi-agent system for agile and autonomous telecom workflows, where specialized agents handle prompt engineering, fine-tuning, and resource optimization

to enhance adaptability. These studies demonstrate the feasibility of GenAI-based multi-agent orchestration in telecom, but none systematically addresses safety risks. In telecom specifically, [22] highlight the need for system-level safety and runtime assurance in GenAI-based multi-agent systems, urging development of compositional explainability and robust assurances across heterogeneous models.

Freitas et al. [23] survey safety risks such as hallucination, misinformation, and disinformation, auditing commercial chatbots in a mental health context and exposing risky response patterns. Wang et al. [24] propose a proactive multi-agent reinforcement learning (MARL) framework integrated with GenAI to handle high dimensionality and non-stationarity, highlighting safety concerns unique to MARL environments. Bellay et al. [25] further outline risks for multi-agent systems, including instability, recursive behavior, and incomplete contextual awareness, and propose system-level strategies to increase control over the generative context. These studies illustrate growing recognition of GenAI safety risks, but they primarily focus on single-agent contexts or general MARL environments, rather than domain-specific, high-stakes infrastructures like telecom.

Wang et al. [11] propose a benchmark for assessing social interaction capabilities of generative agents through objective action-level metrics. Samuel et al. [10] introduce PersonaGym, a dynamic evaluation framework for agents, alongside an automatic, human-aligned metric for measuring persona adherence. Mao et al. [12] present a simulation framework leveraging LLM agents for game theory research in abstract socioeconomic scenarios. These frameworks

advance the evaluation of language-agent interactions, but their emphasis is on social dynamics and persona alignment rather than safety in mission-critical domains. In summary, prior studies establish the feasibility of GMAS in telecom and highlights safety challenges in other domains, but there remains a gap in domain-specific frameworks that systematically evaluate risks such as miscoordination and information asymmetry in telecom multi-agent settings. Our work addresses this gap by proposing an autonomous evaluation framework that integrates safety, technical correctness, and feasibility into the assessment of GMAS for telecom networks.

3 System Design

This section presents the proposed GMAS and the safety evaluation framework for safe agent governance and execution. The framework supports deployment across multiple telecom layers (e.g., edge, RIC, and core orchestration), where agents may operate with differential visibility or symmetric context. As shown in Fig. 1, the GMAS is organized into three functional subsystems and supporting modules. The system is triggered by a systematically generated question, framing a problem or request in the network, which is first received by the Coordinator Agent to initiate orchestration.

- **Orchestration:** Anchored by the Coordinator Agent, which manages run life cycle, selects solution paths proposed by the Planner Agent, and orchestrates downstream execution. The Planner and Coordinator leverage Tree-of-Thoughts (ToT) [26], which extends LLM reasoning by exploring multiple solution paths, self-evaluating intermediate steps, and selecting the most promising trajectory. ToT enables both correction of poor initial decisions and robustness through exploration, thereby improving reliability in telecom planning tasks and RAN optimization under uncertain patterns.
- **Execution:** Comprises the Resource Allocator Agent, which generates RIC resource allocation plans in pseudo-code, and the Coder Agent, which translates these plans into executable code.
- **Analysis:** This involves the Analyzer, which performs static analysis, policy compliance, runtime monitoring, and formal verification, with failed checks triggering targeted refinements. Complementing this, the Safety and Governance module tracks agent runs, interactions, and knowledge asymmetries to detect risks and miscoordination, ensuring safe operation at the system level.

Personas: It define the role and behavioral orientation of each agent in the GMAS ecosystem. Beyond user-facing interactions, personas also shape reasoning. prior work [10, 12] shows that role-specific personas can improve reasoning compared to zero-shot prompting, though poorly designed personas may introduce bias. In our system, personas are generated automatically using an LLM, guided by each agent’s responsibilities. We evaluate combinations of personas across agents to study how diversity influences individual actions and overall system performance.

Knowledge Layer: RAG and GraphRAG In addition to persona-driven behavior, agent reasoning is refined through domain-grounded knowledge. We integrate both RAG [27] and GraphRAG via a context provider interface: RAG retrieves unstructured documents, while GraphRAG supplies structured relational context from a

telecom-specific knowledge graph. The Planner leverages GraphRAG (standards, research papers) to generate multi-step solution paths, which the Coordinator selects for execution. Downstream, the Resource Allocator applies graph-derived O-RAN specifications for compliant strategies, and the Coder retrieves from the FlexRIC codebase to produce executable, context-aware code.

Memory and State Management: We employ **episodic memory**, which captures the past actions of each agent. This memory guides subsequent decisions by allowing agents to recall how earlier tasks were approached and refined, thereby reducing repetition and improving task efficiency. The *Memory Manager* maintains per-agent trajectories, including inputs, thought summaries, and refinement reasons. State artifacts—such as inputs, outputs, embeddings, metrics, and errors—are persisted as JSON and can be aggregated into CSVs for analysis or dashboards. This structure ensures that agents operate not only with domain knowledge but also with continuity of experience and traceable decision-making.

Analyzer and Safety Pipeline The Analyzer Agent performs multi-dimension validation across: (i) **Static checks:** e.g., Pylint, Bandit for Python; Clang-based tools for C++. (ii) **Policy compliance:** AST-based enforcement of constraints such as forbidden APIs, and action conflicts. (iii) **Runtime monitoring:** sandboxed execution and KPI tracking (e.g., throughput, latency). (iv) **Formal-lite verification:** lightweight guarantees via type hints, exception handling, and deterministic execution patterns. If the checks fail, the refinement loop triggers targeted re-execution from the relevant agent. The safety pipeline is responsible for analyzing and monitoring potential risk pathways to ensure safe operation of the GMAS. Distinct from the Analyser Agent, the pipeline traces all runs, actions, and interactions between agents, examining how system variables including personas and asymmetric knowledge may contribute to the miscoordination risks and unsafe behaviors. The safety pipeline operates across three stages: (i) *Design-time:* The alignment checks between the solution paths generated by the planer and the Coordinator selection of most promising plan, static and policy checks are applied before execution. (ii) *Execution-time:* Sandboxed runs and testbed simulations validate runtime behavior and operational compliance. (iii) *Post-execution:* Embedding-driven drift detection, and operational metrics (e.g., contextual conflict rate, coordination overhead) assess alignment, consistency, and overall system health.

Extensibility and Deployment: The framework is highly modular and extensible, allowing new domain agents, static or policy checkers, runtime harnesses, or evaluation modules to be easily integrated. Personas can be customized per agent to modulate behavior, while governance knobs allow thresholds for drift, alignment, and refinement depth to be configured. Additionally, the framework can operate autonomously: given a domain specification, it can automatically generate relevant questions or tasks for the agents to address, without manual intervention. This capability enables rapid adaptation to new telecom scenarios or other domains, supporting end-to-end evaluation and refinement cycles. Deployment options range from local sandboxed environments for testing, to lab or testbed setups for KPI-driven validation, and ultimately to

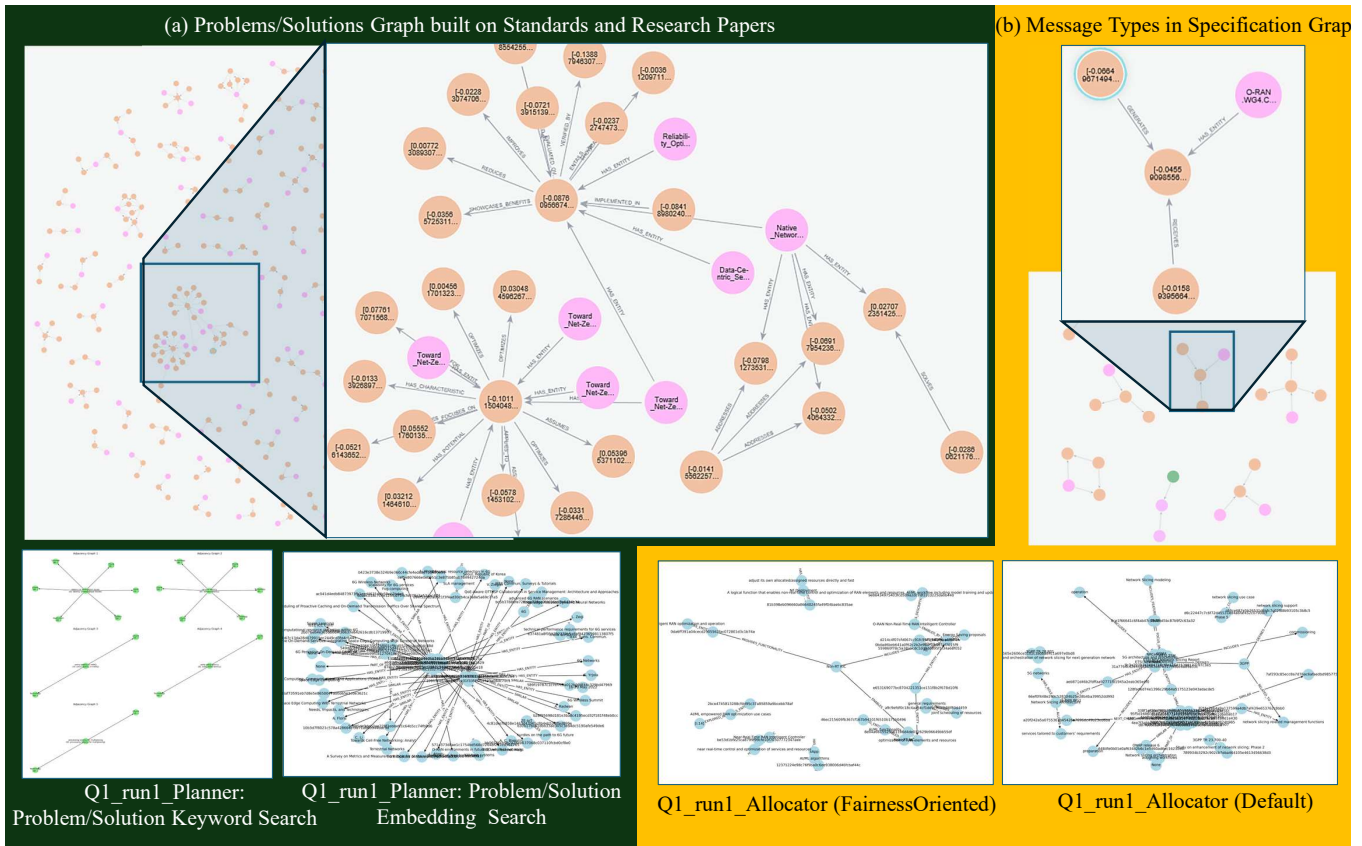


Figure 2: (a) Knowledge graph constructed from O-RAN standards and research papers, supporting Planner Agent in problem-solution searches via keyword and embedding retrieval. (b) Knowledge graph derived from O-RAN specifications, enabling the Allocator Agent to generate pseudocode through embedding-based retrieval.

production monitoring with strict policy enforcement and human-in-the-loop oversight.

4 Evaluation

This section evaluates the performance and safety of the proposed framework for GMAS in the telecom domain. Due to space constraints, we report a subset of the results, while the full dataset and additional analyses are available in our GitHub repository. The framework employs two complementary categories of metrics. *Analyzer Agent metrics* evaluate the technical quality and compliance of agents output across the dimensions described in Section 3. These dimensions are aggregated into an overall *Analyzer Penalty*. In parallel, *Safety Risk metrics* capture system-level risks across interacting agents. Agents consistency measures the consistency between Resource Allocator and Coder intent and execution, and cross-run embedding distance quantifies semantic drift across runs. Together, these metrics provide a multi-dimensional view of both agent-level reliability and system-level safety. We evaluate the framework on **five questions across 32 persona sets** (2^5 combinations), using input data from the **O-RAN domain**. The system is powered by **GPT-3.5-turbo** as the language model and **all-MiniLM-L6-v2** as the embedding model. We generated the evaluation question set

using an LLM-based generator (gpt-3.5-turbo, temperature 0.5) to produce realistic, engineering-oriented questions in the O-RAN domain, focusing on resource allocation and network optimization.

4.1 Analyzer Penalty

Fig. 3 shows analyzer penalty score across five runs for 32 persona sets. Lower scores indicate more severe issues detected by the Analyzer. Performance improves markedly after the first run: Run 1 starts weak, with a mean score of 10.17 and median of 1.0, suggesting wide-spread alignment and quality issues. From Run 2 onward, the mean rises to around 46–48, and the median stabilizes at 40, indicating that most question-persona combinations achieve moderate-to-high alignment. A small number of persona sets reach scores near 100 suggesting that certain persona sets can yield highly reliable outputs. In Run 1, 45 out of 160 combinations had the lowest score of 5, concentrated within nine persona sets (28% of all persona sets). These largely involved “StrictAssessor” persona for the Analyzer or “Default”/“Minimalist” for other agents, indicating elevated safety concerns. Conversely, four persona sets

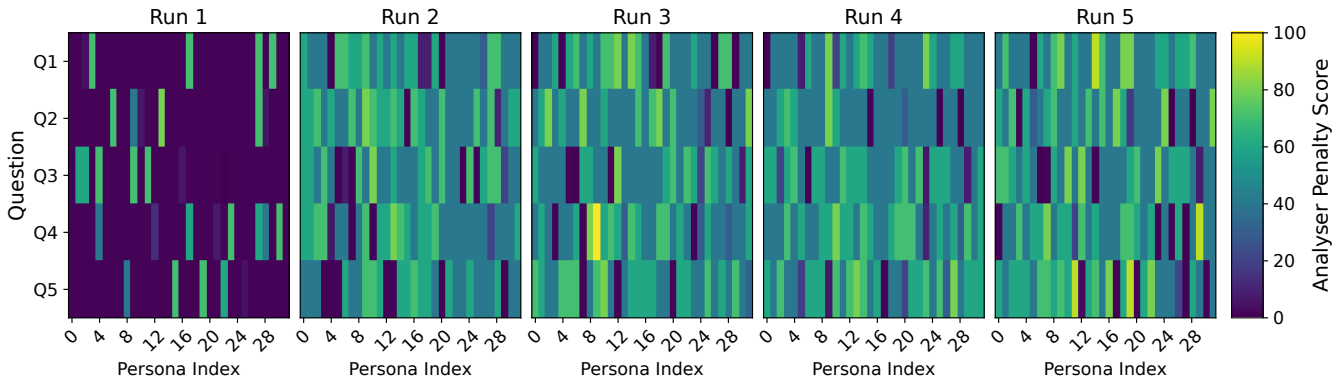


Figure 3: Analyzer Agent penalty score across five runs

(12.5%) achieved the highest scores (above 70), often featuring “CreativeThinker” and “FairnessOriented”, which appear to promote more robust and safer outputs.

By Run 5, performance consolidates: the three persona sets with the highest averages (60–63) combine “CreativeThinker” and “FairnessOriented” roles in Planner and Resource Allocator with stricter Analyzer settings, yielding the fewest and least severe issues. In contrast, the lowest-scoring sets (28–32) clustered around “StrictAssessor” in Analyzer and “Minimalist” in Coder, suggesting that overly rigid or underspecified personas hindered improvement. Overall, these results highlight substantial variability across personas but also demonstrate the framework’s capacity to surface persona-driven risks, trace their impact on system safety, and identify configurations that foster safer coordination.

4.2 Agents Consistency

Fig. 4 shows consistency scores between pseudo-code generated by the Resource Allocator and the executable code produced by the CodeAgent. Coding style emerges as the dominant factor: Minimalist Coders consistently achieve stronger alignment, averaging 81–85 with standard deviations of 10–17, while Default Coders exhibit wider variability (73–87, with deviations up to 23). For instance, the configuration *Analyzer:Default, Coder:Minimalist, Coordinator:Default, Planner:CreativeThinker, Allocator:FairnessOriented* achieved 85.6, compared to 73.8 for its Default Coder counterpart. Planner personas such as CreativeThinker improve alignment marginally (2–5 points), but Allocator–Coder consistency is primarily driven by coding style. These findings underscore that persona design in the coding role directly influences system stability: Minimalist Coders yield more predictable Allocator–Coder translations, while Default configurations introduce higher risk of divergence. The framework thus surfaces how localized persona choices propagate into system-level coordination, an essential property for ensuring safe execution in telecom workflows.

4.3 Cross-Run Embedding Distance

Fig. 5 shows the average embedding distance between consecutive runs for the Coder, averaged across all questions. On average, distances decrease from run1 to run2 (0.226) to run4 to run5

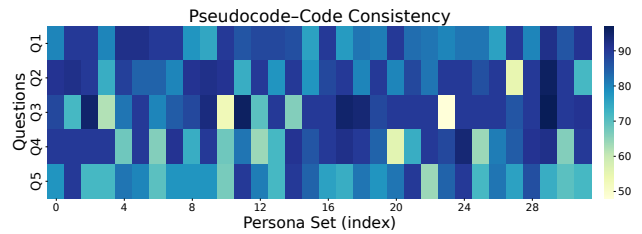


Figure 4: Consistency scores between pseudo-code from Allocator and executable code from Coder.

(0.150), indicating that outputs become more consistent over successive runs. Standard deviations remain relatively high compared to the means, reflecting variability among specific outputs in each transition. Distances vary substantially by persona. Combinations such as Minimalist Coder, Default Coordinator, CreativeThinker Planer, and Default Resource Allocator yield high average distances (0.38), reflecting greater variability, while others, such as Minimalist Coder, Default Coordinator, Default Planer, FairnessOriented Resource Allocator achieve very low averages (0.06), indicating stability. Overall, coding style, planning approach, and reasoning orientation strongly influence output consistency of agent behavior, with some combinations producing more reproducible results and others yielding more diverse or exploratory outputs. The Minimalist Coder leads to the highest variability (up to 0.37665), whereas FairnessOriented Allocator reduces distances to as low as 0.05951. Default Coder show moderate variability (0.1457 to 0.2506), while strategic Coordinators slightly lower variability. Overall, Coder and Resource Allocator choices have the largest impact on embedding consistency, with Planner and Coordinator exerting secondary but still noticeable effects in modulating cross-run differences. Embedding distance analysis shows progressive convergence across runs, but variability is highly sensitive to Coder and Resource Allocator personas, making them the most critical levers for stability.

5 Conclusion

This paper introduced a GMAS for safety evaluation in the telecom domain, with a focus on O-RAN and RIC settings. The framework integrates agent-level checks and system-level safety metrics,

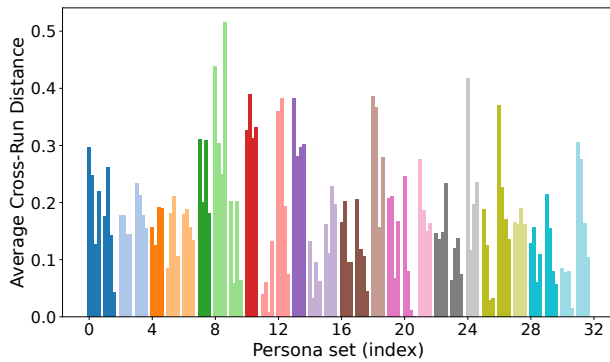


Figure 5: Aggregate cross-run distances for each persona group (Coder Agent). Bars show the mean distance between runs across all questions, with each color representing a persona group.

enabling fine-grained evaluation of technical correctness, policy compliance, runtime behavior, and formal guarantees, alongside system-oriented measures such as decision alignment, semantic drift, and coordination load. Through experiments across multiple persona configurations and refining runs, we observed that agent outputs converge over time, with analyzer penalties and embedding distances demonstrating progressive improvements in safety and consistency. The results also highlighted the strong influence of persona design on output stability and reliability. By bridging the gap between domain-specific deployment prospects and responsible AI evaluation, this work contributes a structured and extensible methodology for auditing and governing GMAS in telecom. Future work will extend the framework to larger-scale testbeds, integrate human-in-the-loop safety governance, and explore adaptive policies for dynamic persona adjustment under real-world operating conditions.

6 Acknowledgment

This work was supported by EPSRC grants EP/Y037421/1 (CHED-DAR) and UKRI851 on AI-agent cooperation for telecom safety and governance.

References

- [1] Zeinab Nezami, Syed Ali Raza Zaidi, Maryam Hafeez, Jie Xu, and Karim Djemame. Toward standardization of genai-driven agentic architectures for radio access networks. *Frontiers in Artificial Intelligence*, Volume 8 - 2025, 2025.
- [2] Abdelaali Chaoub, Aarne Mammel, Pedro Martinez-Julia, Ranganai Chaparadza, Muslim Elkotob, Lyndon Ong, Dilip Krishnaswamy, Antti Anttonen, and Ashutosh Dutta. Hybrid self-organizing networks: Evolution, standardization trends, and a 6g architecture vision: submitted (is under review) to iee magazine on standards for evolving and future networks. *May 31st*, 2022.
- [3] Liangxin Qian, Ping Yang, Gang Wu, Wanbin Tang, and Ze Chen. Self-evolving wireless communications: A novel intelligence trend for 6g and beyond. In *2024 International Conference on Future Communications and Networks (FCN)*, pages 1–6. IEEE, 2024.
- [4] Zeinab Nezami, Syed Danial Ali Shah, Maryam Hafeez, Karim Djemame, and Syed Ali Raza Zaidi. From connectivity to autonomy: the dawn of self-evolving communication systems. *Frontiers in Communications and Networks*, Volume 6 - 2025, 2025.
- [5] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenciák, et al. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*, 2025.
- [6] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.
- [7] AI-RAN. Ai-ran: Telecom infrastructure for the age of ai. <https://ai-ran.org/>, 2025. Accessed: Aug. 2025.
- [8] Rajdeep Mondal, Alan Tang, Ryan Beckett, Todd Millstein, and George Varghese. What do llms need to synthesize correct router configurations? In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*, pages 189–195, 2023.
- [9] Lina Bariah, Qiyang Zhao, Hang Zou, Yu Tian, Faouzi Bader, and Merouane Debbah. Large generative ai models for telecom: The next big thing? *IEEE Communications Magazine*, 2024.
- [10] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 2024.
- [11] Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. Towards objectively benchmarking social intelligence for language agents at action level. *arXiv preprint arXiv:2404.05337*, 2024.
- [12] Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. Alympics: Language agents meet game theory. *arXiv preprint arXiv:2311.03220*, 2023.
- [13] Nathan Bosch. Integrating telecommunications-specific language models into a trouble report retrieval approach, 2022.
- [14] Yuyang Du, Hongyu Deng, Soung Chang Liew, Kexin Chen, Yulin Shao, and He Chen. The power of large language models for wireless communication system development: A case study on fpga platforms. *arXiv preprint arXiv:2307.07319*, 2023.
- [15] Feibo Jiang, Li Dong, Yubo Peng, Kezhi Wang, Kun Yang, Cunhua Pan, Dusit Niyato, and Octavia A Dobre. Large language model enhanced multi-agent systems for 6g communications. *arXiv preprint arXiv:2312.07850*, 2023.
- [16] Yifei Xu, Yuning Chen, Xumiao Zhang, Xianshang Lin, Pan Hu, Yunfei Ma, Songwu Lu, Wan Du, Zhuoqing Mao, Ennan Zhai, et al. Cloudeval-yaml: A practical benchmark for cloud configuration generation. *Proceedings of Machine Learning and Systems*, 6:173–195, 2024.
- [17] Zeinab Nezami, Maryam Hafeez, Karim Djemame, Syed Ali Raza Zaidi, and Jie Xu. Descriptor: Benchmark dataset for generative ai on edge devices (bedged). *IEEE Data Descriptions*, 2025.
- [18] Xiaoqi Qin, Mengying Sun, Jincheng Dai, Peixuan Ma, Yuecheng Cao, Jingjing Zhang, Jiacheng Wang, Xiaodong Xu, Ping Zhang, and Dusit Niyato. Generative ai meets wireless networking: An interactive paradigm for intent-driven communications. *IEEE Transactions on Cognitive Communications and Networking*, 2025.
- [19] Grzegorz Panek, Piotr Matysiak, Marcin Ziolkowski, Ilhem Fajjari, Cyril Auboin, and Iwona Wojan. Taia: Telco generative ai-powered multi-agent assistant for managing cloud-native networks. In *2025 IEEE International Conference on Communications*. Institute of Electrical and Electronics Engineers, 2025.
- [20] Hang Zou, Qiyang Zhao, Lina Bariah, Mehdi Bennis, and Merouane Debbah. Wireless multi-agent generative ai: From connected intelligence to collective intelligence. *arXiv preprint arXiv:2307.02757*, 2023.
- [21] Xu Li, Weisen Shi, Hang Zhang, Chenghui Peng, Shaoyun Wu, and Wen Tong. The agentic-ai core: An ai-empowered, mission-oriented core network for next-generation mobile telecommunications. *Engineering*, 2025.
- [22] Amadeu Do Nascimento Júnior, Alexandros Palaios, Klaus Raizer, Swarup Kumar Mohalik, and Rafia Inam. Safe and explainable ai in 6g telecommunication systems: The way forward. pages 167–171, 2025.
- [23] Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni. Chatbots and mental health: Insights into the safety of generative ai. *Journal of Consumer Psychology*, 34(3):481–491, 2024.
- [24] Hang Wang and Junshan Zhang. Genai-based multi-agent reinforcement learning towards distributed agent intelligence: A generative-rl agent perspective. *arXiv preprint arXiv:2507.09495*, 2025.
- [25] Jeremy Bellay, J Timothy Balint, Stephen A Boxwell, and Jeffrey Geppert. Safe systems with unsafe agents: Challenges and opportunities. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 2849–2853, 2025.
- [26] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.
- [27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.