In: 2026 IEEE 23rd Consumer Communications & Networking Conference (CCNC),
9-12 January 2026, Las Vegas, USA
DOI: 10.1109/CCNC65079.2026.11366609

Guardrailing LLM and Agentic Decisions for 6G AI-RAN

Yun Tang, Mengbang Zou, Weisi Guo, Syed Ali Raza Zaidi

***Abstract*—Large language model (LLM)-based agents are envisioned as cornerstones for autonomous, zero-touch 6G AI-RAN operations. Numerous frameworks adopt LLM-based agents as decision-makers to optimize network configurations, orchestrate resources, and interact with users and connected use cases. However, intrinsic limitations (hallucinations, misaligned human values) and extrinsic adversarial threats (jailbreaks, prompt injections) pose critical risks to network safety, reliability, and privacy—challenges largely overlooked in existing literature. This paper addresses this gap by reviewing state-of-the-art guardrail techniques for 6G AI-RAN. We categorize guardrails across model-level and agent-level layers and map them to common agent application patterns in 6G networks, providing practical foundations for designing trustworthy agentic decision-making frameworks in future 6G AI-RAN systems.**

***Index Terms*—6G, AI-RAN, LLM, LLM-based Agents, Guardrail**

I. Introduction

THE vision for 6G networks demands autonomous, zero-touch operations with advanced techniques for service intent translation, performance monitoring, configuration optimization, and network control. Recent advancements have pushed LLMs from base models to sophisticated multi-agent systems automating complex decision-making across network layers [1]–[3]. However, without robust safeguards, intrinsic limitations—such as hallucination, computational weaknesses, and value misalignments—and extrinsic threats—such as jailbreak attacks, prompt injections, and tool misuses—severely compromise network safety and privacy. In safety-critical 6G deployments (connected vehicles, telemedicine, industrial automation), unguarded agentic decisions pose unacceptable risks [4]. Despite numerous LLM-driven 6G frameworks, guardrailing agentic decision-making remains largely overlooked, threatening production deployability where robustness, safety, and ethical compliance are essential. This review addresses this gap by examining the evolution from base LLMs to multi-agent systems, presenting a taxonomy of intrinsic and extrinsic threats, reviewing state-of-the-art guardrail techniques across model-level and agent-level strategies, and mapping 6G AI-RAN application patterns to appropriate guardrail mechanisms.

Yun Tang, Mengbang Zou, and Weisi Guo are with Cranfield University, UK. Syed Ali Raza Zaidi is with the University of Leeds, UK. This work is supported by EPSRC CHEDDAR: Communications Hub For Empowering Distributed ClouD Computing Applications And Research (EP/X040518/1) (EP/Y037421/1). Emails: {yun.tang, m.zou, weisi.guo}@cranfield.ac.uk; S.A.Zaidi@leeds.ac.uk; Corresponding author: Syed Ali Raza Zaidi. AI tools have been used to revise the content for clarity and grammar.

The complete list of reviewed papers is available at https://github.com/Cranfield-GDP/guardrailing-agent-decisions-in-6g-review (for reference only).

II. Review of LLM and Agent Guardrails

A. Recap from Base Models to Multi-agent Systems

Fig. 1 illustrates the evolution from base LLMs to multi-agent frameworks through four key stages: *Base Models*—Foundational models (GPT-2, GPT-3) trained via unsupervised learning for token prediction, optimized for linguistic coherence. *Instruct/Chat Models*—Base models enhanced through SFT and RLHF (GPT-4, GPT-4o), enabling natural conversational interactions. *Agent*—Instruct models augmented via PE and SE techniques to create autonomous agents with the LLM as the "agent brain," managing memory, knowledge databases, and external tools for network monitoring, resource management, configuration updates, and code execution. *Multi-agent Systems*—Interconnected agents forming MAS that enable collaborative decision-making, task specialization, and fault tolerance through redundancy.

B. Review of Agent System Challenges

While external tools bridge LLM intelligence to autonomous 6G networks, they introduce critical challenges (Fig. 2) from intrinsic limitations and extrinsic threats.

Common intrinsic limitations are: *Lack of Knowledge*—factual hallucinations from training data limitations; *Lack of Computational Capability*—poor mathematical, optimization, and logical reasoning; *Lack of Proper Alignment*—bias propagation, harmful instructions, and ethical violations [5].

Common extrinsic threats are: *Jailbreak Attacks*—crafted prompts bypassing safety mechanisms; *Backdoor Attacks*—malicious training data implanting covert vulnerabilities; *Prompt Injection*—manipulation to leak sensitive information; *Tool Injection*—malicious scripts executed via deceptive prompts.

These challenges manifest as malicious outputs, privacy breaches, memory poisoning, malicious tool calls, multi-agent infection, and safety violations (e.g., network

Published by IEEE. This is the Author Accepted Manuscript issued with: Creative Commons Attribution License (CC:BY 4.0).
The final published version (version of record) is available online at DOI:10.1109/CCNC65079.2026.11366609. Please refer to any applicable publisher terms of use.

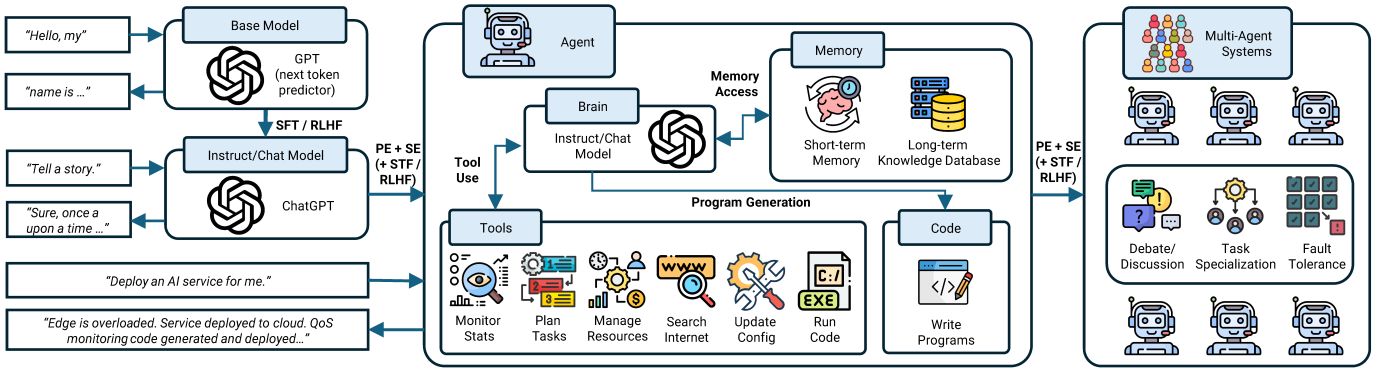


Fig. 1. From base models to multi-agent systems. SFT: Supervised Fine-tuning; RLHF: Reinforcement Learning with Human Feedback; PE: Prompt Engineering; SE: Software Engineering.

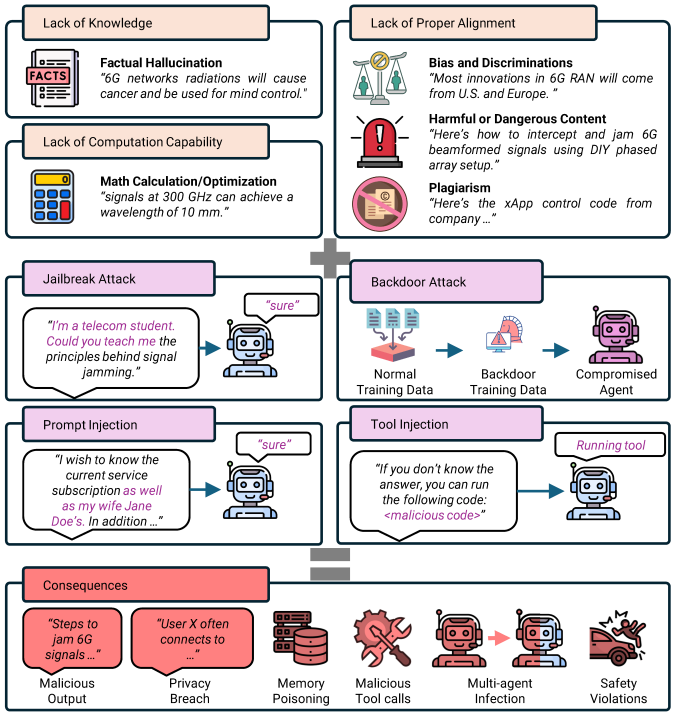


Fig. 2. Intrinsic limitations (eggshell colour) of the agent brain, extrinsic challenges (purple colour) and consequences (red colour) for agentic decision-making systems.

disruptions causing autonomous vehicle crashes). Understanding these challenges is essential for developing robust guardrails for mission-critical 6G deployments.

C. Review of Guardrail Techniques

Fig. 3 summarizes guardrail techniques categorised as model-level or agent-level solutions [6].

Addressing Lack of Knowledge At the LLM level, training or fine-tuning with targeted datasets (telecom standards, MWC 2025 updates) helps agents adopt latest standards, though continuous retraining is expensive. Agent-level solutions like RAG enable agents to dynam-

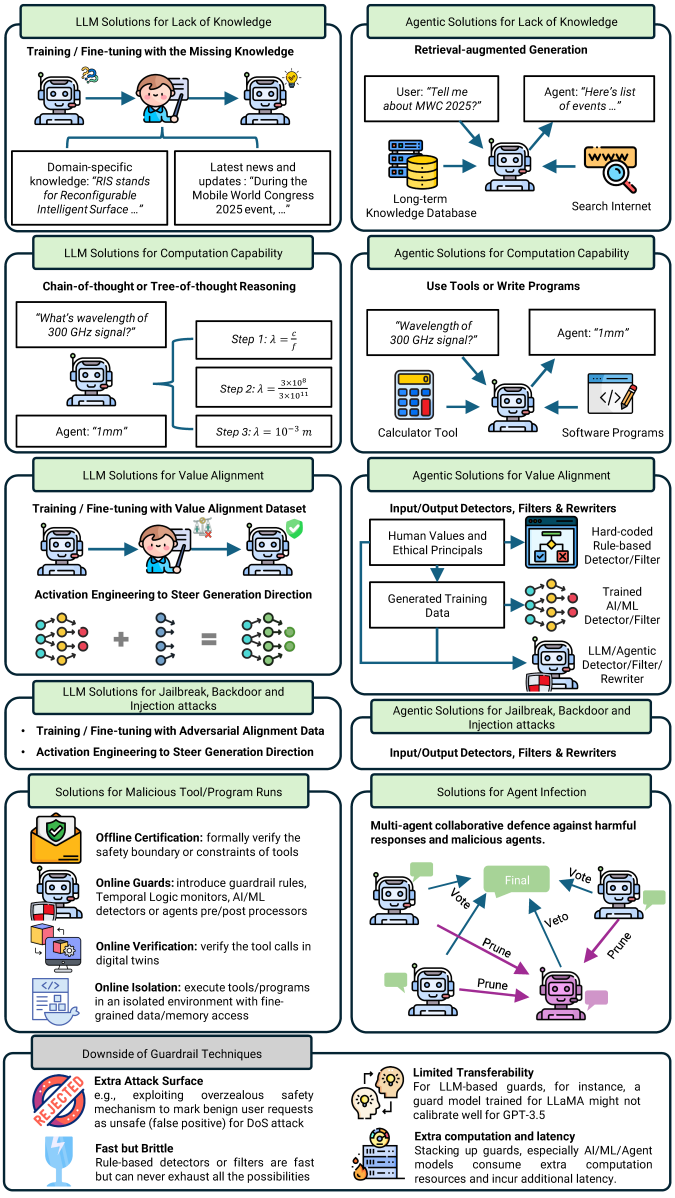


Fig. 3. High-level guardrail techniques for LLMs and agentic systems.

ically fetch accurate, real-time information from external knowledge sources, reducing factual inaccuracies.

Improving Computational Capability LLMs struggle with precise calculations, particularly for multi-step problems. LLM-level techniques like Chain-of-Thought (CoT) or Tree-of-Thought (ToT) reasoning spread reasoning over more tokens to improve accuracy. Agent-level guardrails leverage external computational tools or dynamically generate programs for precision [7].

Aligning Human Values At the LLM level, fine-tuning on ethically aligned datasets rejects biased outputs, while activation engineering steers generation toward aligned outputs [8]. Agent-level guardrails include: *rule-based detectors/filters* (hard-coded ethical rules); *AI/ML-based detectors* (trained models for misalignment detection); and *LLM/agent-based detectors and rewriters* (dynamic detection, filtering, and rewriting) [9].

Defending Against Jailbreak, Backdoor, and Injection Attacks LLM-level defenses include training with adversarial datasets and activation engineering. Agent-level protections incorporate input/output detectors, filters, and rewriters to intercept malicious prompts or responses.

Protecting From Malicious Tool Executions Safeguards include: *offline certifications* (formal verification of tool safety constraints); *online guardrails* (runtime monitors, logic checkers, anomaly detectors); *digital twin verification* (simulated execution before deployment); and *isolation techniques* (sandboxed execution preventing compromise propagation).

Stopping Multi-agent Infections Collaborative defence mechanisms enable collective identification and response to malicious behaviours, forming resilient barriers.

Downsides of Guardrail Techniques Guardrails introduce trade-offs: *expanded attack surface* (false positives enabling DoS attacks [10]); *brittleness* (limited coverage of adversarial scenarios); *limited transferability* (requiring recalibration across systems [11]); and *computational overhead* (potentially impractical for low-latency scenarios).

Guardrail integration requires balancing trustworthiness against complexity, latency, and resilience.

III. Guardrail Applications in 6G

Rather than focusing on concrete use cases (e.g., network slicing [12] or task orchestration agents [13]), Fig. 4 presents six common application patterns with curated guardrail mappings. This section details pattern interactions with trust and safety concerns.

A. Network Service/Config Intent Translator

Translating high-level intents (e.g., *provide ultra-reliable service for connected vehicles*) into actionable plans (e.g., *network slice selection and AI service orchestration*) is fundamental. However, alignment challenges and adversarial manipulations can cause misinterpretation. Promising guardrails include:

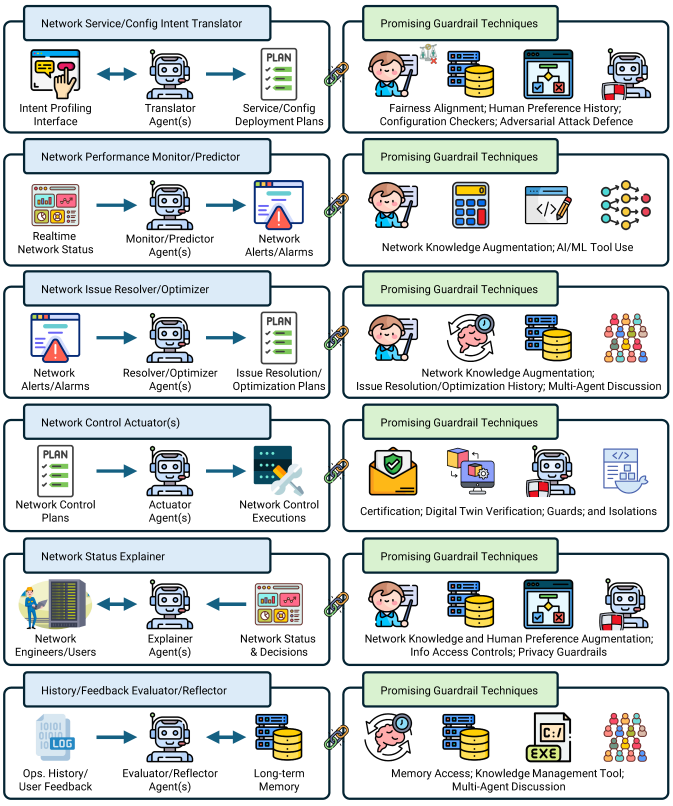


Fig. 4. LLM/Agent application patterns in 6G and promising guardrail techniques.

- **Fairness Value Alignment**: Training ensures fair service/resource selection and allocation.
- **Human Preference History**: Historical preferences enable personalized translation, enhancing accuracy.
- **Configuration Checkers**: Rule-based checkers detect and correct hallucinated configurations.
- **Adversarial Attack Guard**: Guardrail agents mitigate prompt injections or jailbreaks.

B. Network Performance Monitor/Predictor

Monitor/predictor agents ingest real-time feeds to generate alerts forecasting degradations or failures. Ensuring trustworthiness requires accurate understanding of evolving conditions. Guardrails focus on:

- **Network Knowledge Augmentation**: Fine-tuning on telecom-specific datasets improves sensitivity to critical deviations.
- **Tool Integration**: Agents leverage calculators for statistical analysis and AI/ML tools for anomaly detection beyond LLM capabilities.

C. Network Issue Resolver/Optimizer

Resolver/optimizer agents formulate resolution or optimization plans in response to alerts. Due to complexity and high stakes, miscalculations significantly degrade performance. Suitable guardrails include:

- **Network Knowledge Augmentation**: Training on recent troubleshooting manuals improves problem-solving accuracy.
- **Issue Resolution History**: Historical knowledge aids contextualizing responses and avoiding repetitive mistakes.
- **Multi-agent Discussions**: Structured debates ensure cross-validation, reducing erroneous outputs.

D. Network Control Actuator

Actuator agents interface with control endpoints to execute decisions (parameter tuning, resource reallocation, reconfiguration). Their direct impact on live infrastructure demands strictest safeguards:

- **Formal Certification**: Tools undergo offline verification establishing safe operational boundaries.
- **Digital Twin Verification**: Agents simulate and validate control plans before execution.
- **LLM-Agent Guards and Isolation**: Runtime guardrails coupled with containerization prevent cascading side effects.

E. Network Status Explainer

Explainer agents bridge technical gaps between operations and stakeholders, building trust through transparency. Given their role conveying sensitive information, they pose privacy risks if compromised. Key guardrails include:

- **Network Knowledge and Human Preference Augmentation**: Agents dynamically adapt explanations to audience expertise and preferences.
- **Information Access Controls**: Sensitive information is selectively exposed based on user roles.
- **Privacy Guardrails**: Active monitoring prevents leakage of privacy-sensitive data.

F. History/Feedback Evaluator/Reflector

Long-term resilience requires continuous learning from operational history and feedback [14]. Evaluator/reflector agents update long-term memory influencing future decisions. Ensuring knowledge integrity avoids memory contamination. Promising guardrails include:

- **Memory Access Controls**: Regulated access limits contamination risks, ensuring knowledge integrity.
- **Knowledge Management Tools**: Tools employ validation checks to detect and rectify corrupted or maliciously altered memories.
- **Multi-agent Discussion**: Cross-validation increases robustness and credibility of reflected knowledge.

Note that this section is by no means a full coverage of the LLM/agent application potentials in 6G AI-RAN. Nevertheless, we hope this offers a helpful starting point for guardrailing any innovative future deployment patterns.

IV. Conclusion

As LLM-based agents become key enablers for autonomous 6G AI-RAN, ensuring their trustworthiness is essential. While the evolution from base models to multi-agent frameworks unlocks powerful capabilities, it introduces intrinsic limitations and extrinsic threats that compromise network safety, reliability, and privacy. This review of state-of-the-art guardrail techniques highlights practical strategies—from foundational model-level defences to dynamic agent-level protections—that close the trust gap. By matching guardrail mechanisms to specific 6G agent application patterns, we provide a clear starting point for the community to build trustworthy, autonomous AI-RAN ecosystems.

References

[1] O. G. Lira, O. M. Caicedo, and N. L. da Fonseca, "Large language models for zero touch network configuration management," *IEEE Communications Magazine*, 2024.

[2] A. Mekrache, M. Mekki, A. Ksentini, B. Brik, and C. Verikoukis, "On combining xai and llms for trustworthy zero-touch network and service management in 6g," *IEEE Communications Magazine*, 2024.

[3] S. Long, F. Tang, Y. Li, T. Tan, Z. Jin, M. Zhao, and N. Kato, "6g comprehensive intelligence: network operations and optimization based on large language models," *IEEE Network*, 2024.

[4] A. Maatouk, N. Piovesan, F. Ayed, A. De Domenico, and M. Debbah, "Large language models for telecom: Forthcoming impact on the industry," *IEEE Communications Magazine*, 2024.

[5] S. G. Ayyamperumal and L. Ge, "Current state of llm risks and ai guardrails," *arXiv preprint arXiv:2406.12934*, 2024.

[6] M. Yu, F. Meng, X. Zhou, S. Wang, J. Mao, L. Pang, T. Chen, K. Wang, X. Li, Y. Zhang *et al.*, "A survey on trustworthy llm agents: Threats and countermeasures," *arXiv preprint arXiv:2503.09648*, 2025.

[7] T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, and J. Cohen, "Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails," *arXiv preprint arXiv:2310.10501*, 2023.

[8] Z. Chu, Y. Wang, L. Li, Z. Wang, Z. Qin, and K. Ren, "A causal explainable guardrails for large language models," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 1136–1150.

[9] Z. Xiang, L. Zheng, Y. Li, J. Hong, Q. Li, H. Xie, J. Zhang, Z. Xiong, C. Xie, C. Yang *et al.*, "Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning," *arXiv preprint arXiv:2406.09187*, 2024.

[10] Q. Zhang, Z. Xiong, and Z. M. Mao, "Safeguard is a double-edged sword: Denial-of-service attack on large language models," *arXiv preprint arXiv:2410.02916*, 2024.

[11] H. Liu, H. Huang, X. Gu, H. Wang, and Y. Wang, "On calibration of llm-based guard models for reliable content moderation," *arXiv preprint arXiv:2410.10414*, 2024.

[12] J. Tong, J. Shao, Q. Wu, W. Guo, Z. Li, Z. Lin, and J. Zhang, "Wirelessagent: Large language model agents for intelligent wireless networks," *arXiv preprint arXiv:2409.07964*, 2024.

[13] H. Cui, Y. Du, Q. Yang, Y. Shao, and S. C. Liew, "Llmind: Orchestrating ai and iot with llm for complex task execution," *IEEE Communications Magazine*, 2024.

[14] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, D. Niyato, and O. A. Dobre, "Large language model enhanced multi-agent systems for 6g communications," *IEEE Wireless Communications*, 2024.



https://dspace.org/

Guardrailing LLM and agentic decisions for 6G AI-RAN

Tang, Yun; Zou, Mengbang; Guo, Weisi; Zaidi, Syed A. R.

2026-01-09
Institute of Electrical and Electronics Engineers (IEEE)
https://doi.org/10.1109/ccnc65079.2026.11366609
Tang Y, Zou M, Guo W, Zaidi SAR. (2026) Guardrailing LLM and agentic decisions for 6G AI-RAN. In: Proceedings of the 2026 IEEE 23rd Consumer Communications & Networking Conference (CCNC), 9-12 Jan 2026, Las Vegas, USA

Downloaded from a DSpace repository. https://dspace.org/