LLM Signaling for Agentic 6G Network Services

Alexis Solar[†¶], Thierry Lejkin[†], Alexandra Ansiaux[†], Stéphane Rovedakis[¶], Stefano Secci[§¶],
Tomasz Cedzyński[*‡], Halina Tarasiuk[*‡]
[*]Warsaw University of Technology, Warsaw, Poland
[†]Orange Innovation, Châtillon, France
[‡]Orange Poland, Warsaw, Poland
[§]NeuroTel.AI, France
[¶]Cnam, Paris, France
Email: {first-name.last-name}@orange.com; {first-name.last-name}@cnam.fr ; stefano.secci@neurotel.ai

***Abstract*—The emergence of Agentic Artificial Intelligence (A-AI) in future mobile networks raises the need for user equipment (UE) to directly request AI-supported network services from the core network. However, current approaches typically either rely on user plane features to reach agentic functions or require substantial architectural modifications across the RAN and core network to support such interactions. These solutions increase deployment complexity, reduce backward compatibility, and may weaken the efficiency and security properties of existing control-plane procedures. In this paper, we propose a framework for agentic service requests from the UE through the core network control plane, based on limited extensions to non-access stratum (NAS) signaling and reuse of existing N1/N2 procedures. Our approach maintains the security and efficiency benefits of the control plane, and avoids major redesign of network functions. We present the signaling principle, its integration within the 6G Core architecture, and the resulting benefits in terms of deployability, compatibility, and native support for LLM-driven orchestration and automation. We further introduce representative use cases illustrating how agentic services can be accessed directly in the core network through control-plane signaling.**

***Index Terms*—6G core network, agentic AI, LLM orchestration, network automation, intent-based networking**

I. Introduction

Recent progress in Large Language Models (LLMs) has accelerated a shift from conversational artificial intelligence toward systems able to interpret goals, reason over context, invoke tools, and coordinate multi-step actions. In parallel, the telecom community has started to explore how such capabilities can support intent understanding, automation, and service orchestration. More broadly, 6G is increasingly envisioned as an evolution of mobile systems in which AI becomes a native design dimension rather than a purely external add-on [1]–[4]. In this context, AI is no longer only viewed as application-layer traffic carried by the network, but increasingly as an infrastructure service capability that may interact with network control, exposure, and orchestration functions.

For mobile networks, this evolution raises a key architectural question: beyond using AI internally for network optimization and management, how should a user equipment directly request an operator-provided service from the core network? This question is central to future services in which the operator does not only provide connectivity, but also exposes agentic and AI computing capabilities as part of the network service itself. From this perspective, the core network is a natural integration point because it already concentrates key control, policy, and service exposure functions.

This issue is particularly important in 5G Standalone (5G SA), whose core network already provides a structured control-plane architecture for secure and efficient UE-core interactions [5]–[7]. These properties make the control plane a promising candidate for introducing operator-managed A-AI services, provided that the required changes remain limited. However, enabling UE-initiated LLM-based services is not simply a matter of placing an LLM somewhere in the architecture. The key challenge is to define an access and signaling model that remains technically coherent with operator constraints. Existing directions tend either to reach intelligent functions through the user plane, similarly to external application services, or to rely on deeper architectural changes to embed agentic capabilities more natively in future systems. While both directions are relevant, they also involve trade-offs in terms of deployment complexity, backward compatibility, operator control, and preservation of the efficiency and security benefits of current control-plane mechanisms [3], [5], [6]. Guillemin *et al.* [8] examined how 5G core network capabilities can support a practical transition to 6G challenges by preserving backward compatibility with existing 5G interfaces and protocols.

In this paper, we investigate an incremental deployment path, which stays close to current 5G SA principles (in line with [8]), while enabling new 6G service models. We propose a lightweight framework for UE-initiated A-AI service requests through the control plane, based on limited extensions to NAS (Non-Access Stratum) signaling and reuse of existing N1/N2 procedures. The proposed approach provides a pragmatic integration path for operator-managed LLM-driven orchestration and automation without requiring a disruptive redesign of the mobile core.

The main contributions of this paper are as follows:

- We identify the mismatch between emerging UE-initiated agentic service requirements and current UE-to-core interaction mechanisms in mobile networks.
- We introduce a framework for agentic service requests from the UE, based on limited NAS extensions and the

ISBN 978-3-903176-82-9

reuse of existing N1/N2 procedures.

- We propose a multi-agent orchestration framework for agentic service requests, enabling the dynamic coordination of 3GPP functions and external vertical services within the operator network.
- We illustrate how the proposed approach remains aligned with current 5G SA core-network principles through representative use cases, from which we derive key implications, limitations, and open challenges.

The remainder of this paper is organized as follows. Section II reviews related work, analyzes the limitations of existing approaches and highlights the main gap in the current state of the art. Section III presents the proposed framework. Section IV illustrates the framework through representative use cases. Section V discusses open challenges and future perspectives. Section VI concludes the paper.

II. Related Work

The current 5G Core is built around a service-based architecture (SBA), in which core network functions expose services to each other through standardized interfaces. In this architecture, the control plane is structured around network functions providing services through APIs, while signaling toward the UE relies on NAS procedures defined by 3GPP [5]. AI has been introduced in 5G Core Network Architecture since its early beginning in 3GPP Release 15 with the NWDAF (Network Data Analytics Function). Nevertheless, this function aims at provisioning 5GC NFs or Application Functions (AFs) with analytics on UE behavior and Location or Network performances. If those analytics may leverage Artificial Intelligence/Machine Learning (AI/ML) technologies to provide advanced analytics, NWDAF remains a passive function, not taking decisions or triggering actions.

A. Agentic Networked Systems and Natural Language Interfaces

Recent advances in LLMs have enabled the emergence of agentic systems capable of reasoning and interacting with external tools. Such capabilities have recently been explored in networking contexts, notably to support natural language interfaces for intent-based network management. These approaches aim to simplify network control by allowing high-level user intents to be translated into network configurations.

Predictive AI and ML technologies are now quite frequently deployed in network Operation and Maintenance (OAM) systems for anomaly detection or other situational awareness. But those technologies do not give enough autonomy to the network to reach the level 4 of Autonomous Networks (AN) with self-configuration, self-healing, and self-optimization capabilities [9]. To achieve this goal, agents with advanced autonomy are needed. Therefore, quite intensive research efforts are underway in this field led by academics and industrials [10], [11] and standardization, as ETSI proposes the Experience Networked Intelligence (ENI) architecture [12].

Recent research has also begun exploring the integration of those agentic paradigms directly within the architecture of future mobile core networks. Several industrial and academic proposals envision a 6G core where autonomous AI agents assist or orchestrate network functions to enable more adaptive and automated service management. [13], [14] have proposed the concept of an "A-core" (Agentic-AI core), where AI agents interact with core network services to dynamically coordinate network behavior and service provisioning. Intents are at the heart of this architecture, used by customers to request the A-core. These intents are interpreted by Network agents. This A-core is a significant evolution of the 5G-advanced core networks.

3GPP will standardize 6G as a native AI network, encompassing the fulfilment of intents from UEs and AFs, and other AI capabilities and technologies such as potentially AI agents. The study phase for the 6G core definition started in 3GPP since 2025 [15]. Many architectural proposals have been made so far (meeting SA2#173). Even if they can be categorized into user or control plane solutions for the forward of Intents from UEs to 6G core network and if this categorization can be refined according to the type of Intent handler based on Network AI agents or not, the solutions differ on many other points: evolution of the NAS protocol more or less disruptive, NAS endpoint, entity responsible for routing requests based on intent, etc. The work is ongoing, and it is likely that the next meetings will bring other solutions to meet the fulfillment of intent requests based on Network AI agents. These architectural proposals must be largely refined and consolidated.

B. State of the Art Gap

In 5GS, UE and core network are often engaged in complex interactions to achieve the UE services. They use intensively the rich NAS protocol, which nevertheless does not allow for gathering several UE services nor to plan or orchestrate them. UE intent-based requests forward application semantics rather than explicit technical parameters simplifying and alleviating the exchanges between UE and core network, while enriching and paving the way to innovative UE services beyond connectivity. Moreover, in 5GS new services or functionalities deployed in the core network may induce UE-side developments and so quite significant feature development time before being operational. Use of intents reduces this Time To Market by limiting the developments at the core network. The more intelligent the entities handling the intent, the greater the gain when using this new type of interactions between the core and the UE. AI agent technology is an opportunity to maximize this gain. Core network control plane is very secured and resilient serving as the backbone of the operator's network. Intents, concise expressions, forwarded on this plane will profit from this native high level of security and resiliency compared to user plane. NAS protocol is rich enough to readily transport the Intents with limited extensions. Therefore, intents transfer over NAS combined with an Intent handler sub-system based on AI agents seem to offer many perspectives to the operators while following a realistic, backward compatible and smooth path of evolution of the current 5G System towards 6G.

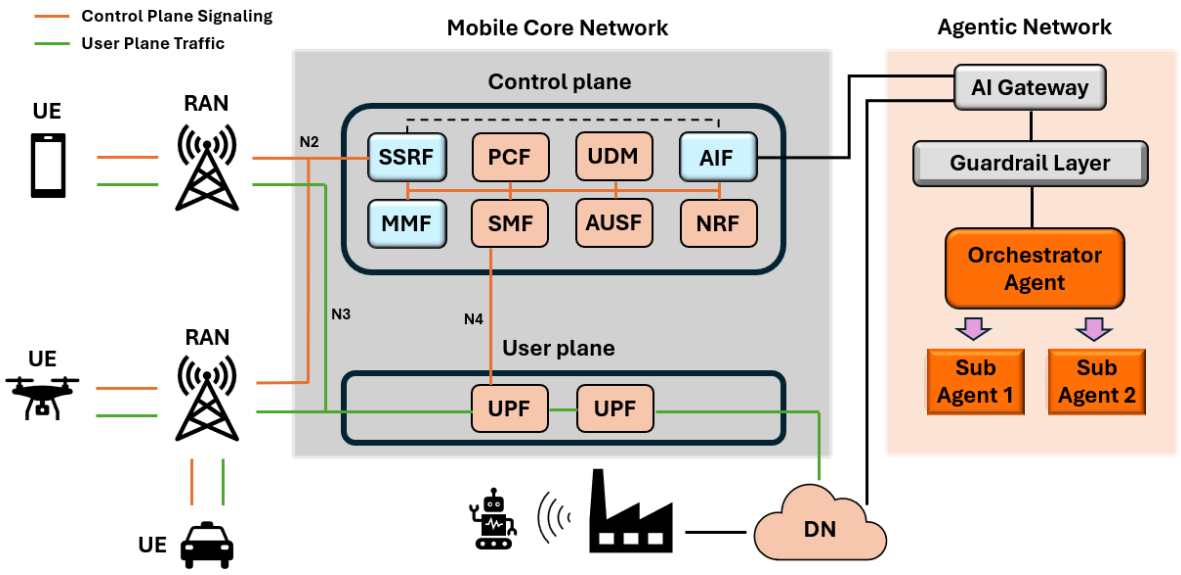

Fig. 1. Architecture for Agentic Service Request handling and orchestration over the 6G core network.

III. Proposed Framework

We first provide the overview of the system, then the control-plane operations and the network design challenges.

A. System Overview

Figure 1 presents the system architecture, consisting of:

- Mobile network: a Radio Access Network (RAN) and 5G Core Network, comprising control plane and user plane functions. Moreover, a dedicated in-network Artificial Intelligence Function (AIF) is introduced in the control plane to handle the agentic service requests. In addition, an agentic service access procedure is defined reusing existing signaling procedures for service request exchange and standard network mechanisms for subsequent data connectivity establishment. We propose splitting the AMF (Access and Mobility Management Function) into two functions; SSRF (Security Session and Routing Function) and MMF (Mobility Management Function), in order to manage this new service in a more flexible way by isolating control-plane operations (managed by AMF initially): SSRF is responsible for access control and for routing agentic service requests to the AIF; meanwhile, the MMF takes over the management of mobility.
- Agentic network: an operator-managed agentic network accessible through a dedicated gateway (AI Gateway), either by the Control Plane through an AIF or by the Data Network (DN). The agentic network is responsible for request interpretation, reasoning and service orchestration.
- Equipment: User equipment (UE) connected either through the RAN or via the DN.

B. Agentic service invocation control plane

The proposed signaling design prioritizes the control plane for agentic service invocation, and uses the user plane only when additional data transfer is required. In particular, short requests such as intents, commands, and concise service interactions are conveyed through existing NAS signaling procedures, thereby avoiding the overhead of systematic user-plane establishment.

To this end, the proposal reuses Uplink and Downlink NAS Transport messages to carry agentic request and response payloads. More specifically, a generic Uplink payload container along with a payload container type are used to encapsulate agentic request and response in the NAS messages [5], [7]. Thus, a lightweight extension to the NAS-transported payload is introduced, in the form of an additional Type-Length-Value (TLV)-based container used to identify agentic-service-related messages. This enables the SSRF to distinguish such messages from conventional NAS-transported information. The agentic requests are forwarded to the AIF via an HTTP/2 service call, and finally the AIF transmits the request to the AI Gateway for execution. If a response needs to be sent to the UE, then the AI Gateway repackages the response into a Downlink payload Container and asks the SSRF to forward it to the UE. In addition, the agentic network may need to send commands to another platform in response to a user request. These commands can either be transmitted via the user plane through the AIF and UPF or via a dedicated link connected to the DN.

Hence, the proposed solution avoids the definition of a new end-to-end signaling framework, is incrementally deployable on top of current 3GPP procedures, hence eases integration into 5G deployments.

Figure 2 illustrates the proposed control-plane signaling procedure, which can be summarized in four phases:

- Phase I – the UE generates an agentic service request and encapsulates it into a payload conveyed through an UL NAS Transport message, together with an additional TLV-based agentic service identifier;

- Phase II – the request is relayed by the radio access and core network signaling path toward the SSRF using existing NAS procedures;
- Phase III – the SSRF detects the agentic-service identifier, extracts the associated payload, and forwards it to the AIF, which relays it to the AI Gateway in the agentic network;
- Phase IV – the agentic network interprets and processes the request, performs reasoning and service orchestration, and optionally returns a response through downlink NAS signaling.

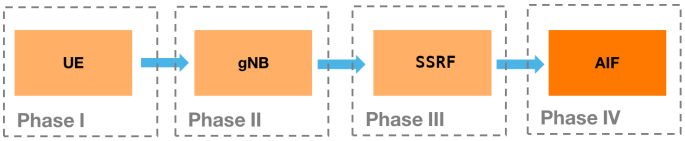

Fig. 2. The extended NAS signaling procedure.

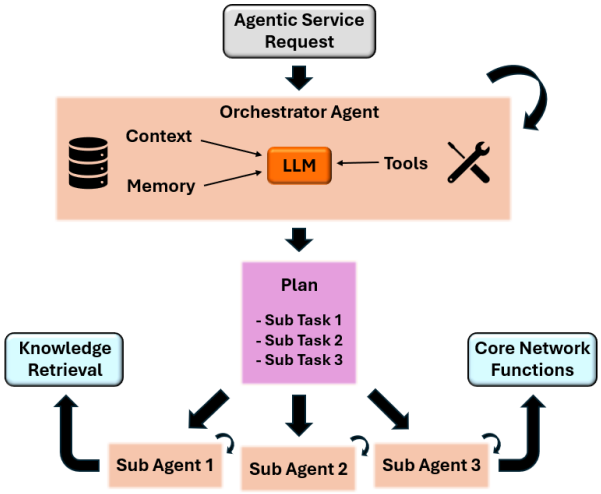

Fig. 3. Architecture of the proposed agentic network.

C. Agentic Network Design

We briefly clarify the agentic processing model considered in this work. The proposed network model is conceived as a stateful agentic system, rather than a standalone LLM, in order to support UE requests that may involve not only queries, but also service commands, multi-step tasks, and context-dependent orchestration.

1) Agentic processing model: Unlike a conventional LLM, which maps an input prompt to a textual output, an LLM-based agent combines language generation with context handling, memory, tool usage, and action selection [16]. Recent work has shown that coupling reasoning and action improves task completion in interactive environments, while tool use extends the model beyond its parametric knowledge [17], [18]. In our case, this is essential because operator-provided agentic services must process short UE intents or commands as part of an evolving task rather than as isolated prompts.

Accordingly, the agentic network maintains a task state that may include the current request, interaction history, service context, execution status, and operator or enterprise constraints. It can rely on short-term or persistent memory, and invoke external tools such as vertical-service connectors, retrieval modules, verification components, or deterministic service adapters. Depending on the situation, the system may return a response to the UE, request clarification, invoke a specialized capability, or trigger a validated service action.

2) Internal organization of the multi-agent system: Figure 3 illustrates the logical organization of the proposed agentic network.

The proposed agentic network follows a centralized orchestration model with specialized sub-agents [19], [20]. After request normalization and context association, an orchestrator agent determines whether the request can be handled directly or should be decomposed into subtasks delegated to specialized components. This organization is motivated by the heterogeneity of the target services, which may involve network-related decisions, vertical-service semantics, context retrieval, or action verification.

Specialized sub-agents can therefore be designed around bounded capabilities and explicit interfaces. For instance, one component may prepare network-related actions, another may encapsulate vertical-service logic such as robot or drone control, and another may retrieve context or verify consistency before execution. The orchestrator remains responsible for task decomposition, sequencing, and aggregation of intermediate results. This structure improves modularity, traceability, and extensibility, since new services can be integrated through additional agents or adapters without redesigning the overall control logic.

3) Semantic guardrails and action validation: In the proposed architecture, conventional mobile-network security and agentic safety address different threat surfaces. The 5G SA control plane already provides strong mechanisms for authentication, integrity protection, and secure signaling transport [5], [21]. However, these mechanisms do not by themselves mitigate failures specific to LLM-based and agentic systems, such as prompt injection, unsafe tool use, conflicting instructions, or semantically invalid actions [22], [23]. These risks stem from the transformation of natural-language requests into executable operations, which introduces a semantic attack surface distinct from classical network threats.

For this reason, the proposed agentic network incorporates semantic guardrails around the orchestration process [24]. These controls include input-level filtering to detect ambiguous or adversarial instructions, policy-grounded decision constraints derived from user rights and operational rules, and pre-execution validation to ensure that any selected action is authorized, coherent, and safe before external tools or services are invoked [25]. In this sense, network security protects the signaling path, while semantic guardrails protect the interpretation and operationalization of the request. The two layers are therefore complementary.

IV. Illustrative Use Cases

A. Use Case 1: Remote Industrial Robot Assistance

Consider a factory scenario in which an operator remotely asks an industrial robot to carry out a task, such as moving to a target area and performing a given action. Here, the proposed framework allows the request to be interpreted and coordinated by an agentic platform that supervises execution and supports interaction with the robot throughout the service. The scenario may also include optional monitoring functions, such as live video feedback, to assist the operator while the task is underway.

Service Workflow

1) The operator sends a short natural-language command from the UE, for instance asking the robot to move to a target area and perform a given action. The request is conveyed in an UL NAS Transport message together with the agentic service identifier.
2) After crossing the RAN, the message reaches the core network control plane, where the SSRF detects the agentic service container, extracts the payload, and forwards it to the AIF.
3) The AIF passes the request to the AI Gateway. At this stage, the guardrail layer checks whether the request complies with authorization rules, enterprise policies, and operational safety constraints. If necessary, the UE may receive a clarification request or a rejection through downlink NAS signaling.
4) The orchestrator agent then interprets the command, evaluates task feasibility and robot status through dedicated tools or sub-agents, and converts the high-level intent into structured instructions for the industrial control domain.
5) These instructions are delivered to the robot-control system, which initiates the requested operation. At the same time, the agentic platform may return an acknowledgment or progress update to the UE over the control plane.
6) When live monitoring is needed, a video session is set up through the conventional user plane via the DN path. Temporary QoS adaptation may be applied to support reliable real-time feedback during execution.
7) The platform may also request additional communication support for this monitoring phase, for example an adapted QoS treatment for the video stream. This preserves the separation between concise service control over the control plane and bulk media transport over the user plane.
8) After the task ends, the platform notifies the UE of completion and releases any temporary service-specific resources or QoS adjustments.

B. Use Case 2: Urban Autonomous Drone Mission Adaptation

Consider an urban scenario in which an autonomous drone, connected to the public 6G network through a SIM/eSIM, carries out a mission such as infrastructure inspection, parcel delivery, or area monitoring. In contrast to the industrial robot case, the drone evolves over a large public area and must cope with changing mobility, connectivity, and regulatory conditions. The proposed framework allows the drone itself to request network-assisted agentic support whenever an unforeseen event arises during the mission, such as a temporary restricted zone, degraded radio conditions, or a low-battery situation requiring trajectory adaptation.

Service Workflow

1) While in operation, the drone detects a condition that requires mission adaptation and sends a short request in an UL NAS Transport message together with the agentic service identifier.
2) Using the same control-plane processing principle as in Use Case 1, the message is forwarded to the AIF and then to the AI Gateway, where the guardrail layer checks the drone identity, mission authorization, and relevant urban safety constraints.
3) The orchestrator agent then collects the information needed to assess the situation, including mission state, location, battery level, radio conditions, and external data related to airspace restrictions.
4) From this context, the agentic platform derives an updated mission strategy, for example by computing a safer route, selecting a temporary waiting area, identifying a fallback landing point, or requesting temporary communication support for a critical phase.
5) The selected decision is sent back to the drone through downlink NAS signaling as concise mission-level instructions. When continuous telemetry, video streaming, or remote supervision is needed, a conventional user-plane session may also be established or adjusted with suitable QoS treatment.
6) Once the situation has been handled or the mission has ended, the platform reports the outcome and releases any temporary communication or service-specific resources.

V. Open Challenges and Discussion

Several challenges remain before such a framework can be considered mature for deployment in future mobile networks.

First, experimental evidence on UE-initiated agentic service signaling through the mobile core is still limited. Although the proposed approach builds on existing control-plane principles, standardized NAS procedures, and an incremental integration path, the underlying agentic technologies remain recent. Their integration into telecom environments therefore raises open questions regarding scalability, latency, robustness, and compliance with operator-grade constraints. Further experimentation is needed to determine whether such a framework can operate at scale while preserving the reliability and predictability expected from mobile-core services.

Second, the broader architectural direction of the future core network remains open. Current visions range from deeply integrated agentic-core designs to more conservative approaches

that stay close to 5G SA principles and rely on external or loosely coupled agentic functions. A fully agentic core may provide greater flexibility, but it also entails higher migration costs and stronger operational uncertainty. Conversely, overly conservative designs may preserve deployability, but at the cost of limiting flexibility and slowing architectural evolution. The proposed framework is intended as a middle path between these two extremes.

Third, the practical deployability of the proposed approach also depends on UE-side access to signaling capabilities. In current mobile systems, NAS signaling remains largely invisible to applications running on the UE, which prevents over-the-top services from directly issuing agentic commands through the control plane. As a result, the proposed framework would require device and platform vendors to support the relevant signaling interaction mechanisms on the UE side. Since this dependency lies largely outside the operator domain, it remains an important uncertainty for real-world adoption.

Finally, agentic service requests involving larger payloads or continuous exchange cannot be transmitted via the control plane. In fact, this implies the transmission of a higher volume of traffic than the control plane was designed for. Therefore, the standard user-plane connectivity can be set up through conventional network procedures and specific routing rules on UPFs can be configured to forward these agentic service requests to the AI Gateway. In this situation, the goal is to provide a secure connection through the operator's infrastructure to route this traffic to the agentic network without necessarily going through the public network.

VI. Conclusion

This paper proposed a lightweight framework for supporting agentic service requests from the UE in future mobile networks. By relying on limited NAS extensions and the reuse of existing control-plane procedures, the proposed approach offers an incremental path toward integrating agentic capabilities provided by the operator into the 6G core without requiring a disruptive architectural redesign. The presented architecture and use cases show the relevance of this direction for future AI-native network services. At the same time, important questions remain open, in particular regarding scalability, safety, standardization, and practical deployment. These aspects will be addressed in future work through prototype implementation and experimental evaluation.

Acknowledgements

This work was partially funded by BPI France contracts nb DOS0275875/00 and DOS0239248/00.

References

[1] J. Hong, N. V. Tu, and J. W. Hong, "A comprehensive survey on llm-based network management and operations," *International Journal of Network Management*, vol. 35, no. 6, 2025.

[2] ITU-R, "Framework and overall objectives of the future development of imt for 2030 and beyond," International Telecommunication Union, Recommendation ITU-R M.2160-0, 2023.

[3] Ericsson, "Ai agents in the telecommunication network architecture," White Paper, 2025.

[4] 3GPP, "Functional architecture and information flows for aiml enablement service," 3GPP TS 23.482, release 19, under change control.

[5] ——, "System architecture for the 5g system (5gs)," 3GPP TS 23.501, under change control.

[6] ——, "Procedures for the 5g system (5gs)," 3GPP TS 23.502, under change control.

[7] ——, "Non-access-stratum (nas) protocol for 5g system (5gs); stage 3," 3GPP TS 24.501, under change control.

[8] F. Guillemin, X. Huang, and S. Lataste, "5g evolutions for a smooth migration to 6g: The way forward for the core network," in *4th International Conference on 6G Networking (6GNet)*.

[9] TM Forum, "Autonomous Network Levels Evaluation Methodology," TM Forum, Introductory Guide IG1252, Jun. 2023.

[10] K. J. K. Feng, D. W. McDonald, and A. X. Zhang, "Levels of autonomy for AI agents," *CoRR*, vol. abs/2506.12469, 2025.

[11] D. B. Acharya, K. Kuppan, and D. Bhaskaracharya, "Agentic AI: autonomous intelligence for complex goals - A comprehensive survey," *IEEE Access*, vol. 13, pp. 18 912–18 936, 2025.

[12] ETSI, "Experiential Networked Intelligence (ENI); System Architecture," ETSI, Tech. Rep. GS ENI 005 V4.1.1, Jan. 2026.

[13] X. Li, W. Shi, H. Zhang, C. Peng, S. Wu, and W. Tong, "The agentic-ai core: An ai-empowered, mission-oriented core network for next-generation mobile telecommunications," *Engineering*, 2025.

[14] W. Tong, W. Huo, T. Lejkin, J. Penhoat, C. Peng, C. Pereira, F. Wang, S. Wu, L. Yang, and Y. Shi, "A-core: A novel framework of agentic ai in the 6g core network," in *International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2025, pp. 1104–1109.

[15] 3GPP, "Release 20," [Online]. Available: https://www.3gpp.org/specifications-technologies/releases/release-20, [Accessed: Apr. 8, 2026].

[16] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[17] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," 2023.

[18] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *Advances in neural information processing systems*, vol. 36, pp. 68 539–68 551, 2023.

[19] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," *arXiv preprint arXiv:2402.01680*, 2024.

[20] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin *et al.*, "Metagpt: Meta programming for a multi-agent collaborative framework," in *The 12th Int. Conference on Learning Representations (ICLR)*, 2023.

[21] 3rd Generation Partnership Project (3GPP), "Technical Specification Group Services and System Aspects; Security architecture and procedures for 5G System (TS 33.501)," 3GPP, Tech. Rep. TS 33.501, 2026.

[22] Y. Liu, G. Deng, Y. Li, K. Wang, Z. Wang, X. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng *et al.*, "Prompt injection attack against llm-integrated applications," *arXiv preprint arXiv:2306.05499*, 2023.

[23] M. A. Ferrag, N. Tihanyi, D. Hamouda, L. Maglaras, A. Lakas, and M. Debbah, "From prompt injections to protocol exploits: Threats in llm-powered ai agents workflows," *ICT Express*, 2025.

[24] M. Shamsujjoha, Q. Lu, D. Zhao, and L. Zhu, "Swiss cheese model for ai safety: A taxonomy and reference architecture for multi-layered guardrails of foundation model based agents," in *22nd International Conference on Software Architecture (ICSA)*. IEEE, 2025, pp. 37–48.

[25] 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.