AURA: Adaptive Unified Reasoning and Automation with LLM-Guided MARL for NextG Cellular Networks

Narjes Nourzad
Department of Electrical and Computer Engineering
University of Southern California
Los Angeles, CA 90007
nourzad@usc.edu

Mingyu Zong
Department of Computer Science
University of Southern California
Los Angeles, CA 90007
mzong@usc.edu

Bhaskar Krishnamachari
Department of Electrical and Computer Engineering
Department of Computer Science
University of Southern California
Los Angeles, CA 90007
bkrishna@usc.edu

Abstract

Next-generation (NextG) cellular networks are expected to manage dynamic traffic while sustaining high performance. Large language models (LLMs) provide strategic reasoning for 6G planning, but their computational cost and latency limit real-time use. Multi-agent reinforcement learning (MARL) supports localized adaptation, yet coordination at scale remains challenging. We present AURA, a framework that integrates cloud-based LLMs for high-level planning with base stations modeled as MARL agents for local decision-making. The LLM generates objectives and subgoals from its understanding of the environment and reasoning capabilities, while agents at base stations execute these objectives autonomously, guided by a trust mechanism that balances local learning with external input. To reduce latency, AURA employs batched communication so that agents update the LLM's view of the environment and receive improved feedback. In a simulated 6G scenario, AURA improves resilience, reducing dropped handoff requests by more than half under normal and high traffic and lowering system failures. Agents use LLM input in fewer than 60% of cases, showing that guidance augments rather than replaces local adaptability, thereby mitigating latency and hallucination risks. These results highlight the promise of combining LLM reasoning with MARL adaptability for scalable, real-time NextG network management.

1 Introduction

6G cellular networks are expected to provide high data transmission speeds and seamless connectivity, supporting devices from autonomous vehicles to personal gadgets at large-scale events [Banafaa et al., 2023, Chataut et al., 2024]. These advancements have the potential to redefine how we connect and communicate, but this promise has yet to translate into practical, deployable solutions [Maduranga et al., 2024, Shahjalal et al., 2023, Cui et al., 2024]. Next-generation networks face increasing challenges in managing dynamic and unpredictable traffic. High user density, coupled with ultra-low latency demands for applications such as holographic imaging and haptic communication, requires

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

sophisticated algorithms [Maduranga et al., 2024, Dogra et al., 2023]. Consequently, the imperative to balance reliability, energy efficiency, high data rates, and low latency has led researchers to envision 6G as AI-driven networks. AI integration aims to improve throughput, reliability, and latency, while enabling self-optimizing operations through intelligent resource allocation and adaptive traffic management [Yang et al., 2020, Noman et al., 2023].

Traditional heuristic methods fall short in meeting these demands due to their limited scalability and difficulty in quantifying performance gaps [Abasi et al., 2024], prompting a shift toward machine learning (ML) methods [Kim et al., 2023, Wang et al., 2023b]. ML and deep learning (DL) algorithms have considerably improved network performance, but they face major challenges in real-world 6G settings, because of their reliance on annotated data. The NP-hard nature of labeling makes supervised learning impractical for dynamic settings, lacking both scalability and adaptability [Cui et al., 2024]. Reinforcement learning (RL) offers an alternative to supervised methods by eliminating the need for labeled data. Multi-agent reinforcement learning (MARL) extends RL to environments with multiple agents that must coordinate or compete [Bhati et al., 2023]. By enabling localized decision-making, MARL provides scalability and robustness [Sun et al., 2023], making it a natural fit for 6G network management [Chu et al., 2019]. However, current MARL algorithms struggle to learn distributed policies for cooperative tasks, particularly in sparse-reward, dynamic environments with large action spaces, characteristics of next-generation networks [Feriani and Hossain, 2021].

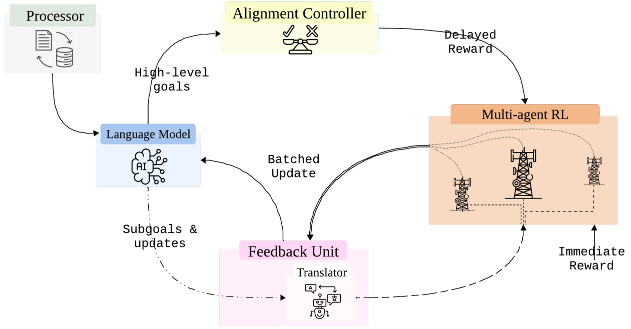


Figure 1: Illustration of the *AURA* Architecture. A cloud-based LLM sets high-level objectives from unified inputs processed by the multimodal encoder. Distributed MARL agents adapt locally through actionable subgoals, while the aliment controller aligns policies and assigns rewards. Iterative feedback refines decisions for dynamic network management.

Centralized training with decentralized execution (CTDE) is widely adopted to overcome the limitations of independent learning [Wang et al., 2023a]. Nevertheless, CTDE itself faces constraints such as limited agent communication, difficulties adapting to non-stationary environments, and scalability issues as the number of agents grows [Oroojlooy and Hajinezhad, 2023]. Additionally, many MARL-based solutions suffer from high complexity, as agents must encode large amounts of information into their policies. Because these policies are rarely generalizable, they are often trained from scratch, further increasing computational costs and slowing convergence [Yang et al., 2025]. These limitations suggest the need for alternative approaches that combine high-level reasoning with localized adaptability to ensure cohesive multi-agent coordination. In parallel, large language models (LLMs) with billions of pre-trained parameters have demonstrated remarkable capabilities in reasoning, planning, and structured decision-making [Xi et al., 2025]. Their ability to generalize across tasks makes them a compelling tool for network intelligence. In particular, LLMs can function as high-level semantic planners, leveraging in-context learning and prior knowledge [Ahn et al., 2022]. Through planning, LLMs decompose high-level goals into actionable low-level tasks; through reasoning, they structure complex problems into priors and beliefs. However, directly deploying LLMs for fine-grained, real-time network parameter adaptation remains infeasible due to their computational overhead and latency. To reconcile these complementary strengths and weaknesses, we propose *AURA: Adaptive Unified Reasoning and Automation* for NextG cellular networks.

AURA is a hierarchical framework that combines the predictive capabilities of Large Language Models (LLMs) with the real-time adaptability of multi-agent reinforcement learning (MARL). It addresses core 6G challenges such as dynamic traffic patterns, fluctuating user demands, and evolving network conditions. A multimodal processor unifies diverse data sources, which the LLM uses to set high-level objectives and subgoals through techniques like Chain-of-Thought reasoning [Wei et al., 2022] and pre-trained policies. Each base station acts as a MARL agent that autonomously executes its objectives under a trust mechanism, balancing local learning with strategic oversight. A centralized controller coordinates these local actions with global goals through structured rewards. Lightweight communication and batched feedback further enhance scalability and responsiveness, drawing inspiration from reinforcement learning from AI feedback (RLAIF) [Bai et al., 2022, Lee et al.]. We evaluate AURA in a 6G networks operating in a scenario with dynamic user demands. The LLM anticipates congestion and generates preemptive objectives, while MARL agents adapt in real time, taking actions adjusting transmit power and enabling handoffs. Preliminary results show that AURA significantly reduce dropped requests and system failures compared to pure MARL baselines, with the largest gains under normal and high traffic. Notably, AURA achieves these improvements while relying on LLM input in fewer than 60% of cases. This is desirable because frequent LLM queries introduce latency, create a central point of failure, and risk propagating hallucinated or suboptimal suggestions. Since LLMs cannot directly interact with the environment or receive real-time feedback, overreliance would limit adaptability.

2 Methodology

We now introduce *AURA*, an Adaptive Unified Reasoning and Automation framework. Our design is guided by two key objectives: leveraging LLMs for high-level reasoning and ensuring computationally efficient, real-time decision-making via MARL. The following subsections outline *AURA*'s model structure (see Figure 1) and its role in AI-driven network management.

Multimodal Processor for Encoding. The multimodal processor at the LLM input integrates diverse data from modern networks, including time-series metrics, statistical logs, and external sources such as social media, GIS data, and historical traffic. GIS inputs, for example, capture spatial patterns like node distribution or congestion. Since standard LMs are limited to text, the processor transforms all heterogeneous inputs into a shared latent representation aligned with language tokens, preserving numerical, spatial, and textual information for LLM analysis.

LLM for High-Level Planning. The centralized LLM integrates processed representations to predict demand surges and guide both immediate resource allocation and long-term planning. It follows a two-tiered strategy: for common scenarios (e.g., typical congestion), it selects from a repository of pre-trained policies ("offline playbook"); for novel or complex conditions, it applies semantic reasoning (e.g., Chain-of-Thought [Wei et al., 2022]) to adjust reward structures or environment parameters for MARL adaptation. Rather than micromanaging interactions, the LLM defines dynamic subgoals and rewards, enabling MARL agents to implement responses locally. This delegation ensures that global objectives are realized through policies that remain flexible and adaptive at the local level [Zhuang et al., 2024].

MARL Agents for Local Adaptation. At individual base stations, MARL agents adjust parameters (e.g., power) using real-time environmental feedback. They refine policies through trial-and-error and memory-based adaptation, exercising autonomy rather than serving as passive executors of LLM guidance. After each cycle, agents report local actions, conditions, and outcomes, enabling the LLM to update its understanding. Final action choices remain agent-driven and are moderated by a trust mechanism: agents weigh their own policies against LLM suggestions, increasing trust when external input improves performance and decreasing it otherwise. This balance preserves independence while leveraging strategic guidance when beneficial. Agents also receive dual rewards:

(i) *Immediate feedback* from the environment in response to their local actions.

(ii) *Delayed feedback* from the *Centralized Alignment Controller* for their contributions to broader strategic goals.

This hierarchical reward mechanism ensures that the objectives of the individual agent remain aligned with the global network objectives while maintaining local autonomy in decision-making. This process is outlined in Algorithm 1.

Feedback Unit for Communication. To reduce latency in LLM–agent interaction, AURA employs batched communication: subgoal updates and feedback are aggregated at fixed intervals, lowering overhead while preserving responsiveness [Zhuang et al., 2024]. This limits continuous feedback during training, enabling agents to adapt locally while remaining aligned with global goals. To further improve adaptability, AURA incorporates Reflexion-inspired verbal feedback [Shinn et al., 2023]. These lightweight updates provide corrective cues (e.g., reprioritizing subgoals during congestion) without costly retraining, guiding agents to switch strategies under high demand. A *language-to-policy translator* bridges verbal instructions and MARL execution, using semantic parsing (inspired by GPT-like models [OpenAI, 2024]) to convert guidance, such as adjusting exploration or prioritizing goals, into actionable parameters that shape agents' decision-making.

3 Evaluation Scenario: Managing Network Overload Over Varying Traffic Conditions

We evaluate AURA in a custom simulation with two base stations: a *rural* station ($43 - 46$ dBm, max 50 users) and an *urban* station ($30 - 37$ dBm, max 30 users). Users arrive and depart dynamically, each assigned a random signal strength (-120 to -50 dBm) and SNR ($0 - 30$ dB). Episodes begin with randomized power levels, user counts, and channel conditions to ensure variability. Derived policies are used by frameworks as starting point during the testing phase. The evaluation includes three configurations: MARL-ONLY, where agents operate independently without external input; GUIDED MARL, where agents receive high-level suggestions from the LLM and selectively adopt them based on a learned trust mechanism; and AURA, which combines LLM guidance with delayed reward shaping to align local agent behavior with global objectives. Additional experiment details can be found in Appendix C.

4 Experimental Results

We present preliminary results demonstrating the benefits of incorporating different levels of LLM guidance for adaptive network management under varying traffic conditions (Figure 2). Both GUIDED MARL and AURA consistently reduce the number of dropped requests relative to MARL-ONLY, with the largest improvements under normal and high traffic (Table 2a). As illustrated in Table 2b significant difference[1] was observed for both agents under the normal and high traffic conditions when comparing number of dropped requests across all configurations (no significant differences were detected in low traffic). To further identify which methods drive these differences, we apply pairwise Dunn post-hoc tests against the MARL-ONLY baseline (Table 2c – stars denote threshold). These results confirm that AURA yields the strongest improvements ($p < 0.01$), while GUIDED MARL also outperforms the baseline, albeit with weaker evidence. Taken together, these findings indicate that LLM guidance is most beneficial when the network is stressed, where MARL-only agents struggle to sustain reliable handoffs.

Figure 2d further shows that agents rely on LLM suggestions only moderately (below 60%), even in high-traffic scenarios, demonstrating that guidance does not lead to overdependence. AURA exhibits slightly lower reliance than GUIDED MARL, as delayed reward shaping provides additional signals that encourage agents to refine their policies more independently. This highlights that the framework can exploit LLM strategies while preserving local adaptability, thereby reducing latency from frequent queries, avoiding overdependence on a single centralized model, and mitigating errors from hallucinated or non-grounded guidance. Finally, Figure 2e reports system failures, defined as the number of testing steps in which the system failed to serving all requests, directly reflecting the frequency of system breakdowns. As expected, failure rates escalate with traffic intensity, but both GUIDED MARL and AURA substantially reduced these occurrences compared to MARL-ONLY, with AURA achieving the lowest failure counts under high traffic.

5 Conclusion and Future Work

This work introduced AURA, an LLM-guided MARL framework that balances strategic planning with localized adaptability in dynamic cellular networks. By combining LLM objectives with trust-gated action adoption and delayed reward shaping, the framework reduces failures and improves resilience under different traffic dynamics. Preliminary results show reduces failures and dropped requests under stress while limiting LLM reliance to <60%. Future work includes extending the multimodal processor and LLM-based planner to provide richer context and strategic guidance. Key directions are scaling AURA to larger multi-agent deployments, improving robustness to imperfect LLM input through uncertainty estimation and trust calibration, enhancing explainability to support operator oversight, and benchmarking against emerging 6G baselines for realistic evaluation.

[1] Since the normality assumption did not hold, we applied the non-parametric Kruskal–Wallis [Kruskal and Wallis, 1952] test to compare number of dropped calls across all configurations for each agent, followed by pairwise Dunn's test with Holm correction [Dinno, 2015] to identify which configurations differed significantly.

Traffic	Method	Agent$_1$(urban)	Agent$_2$(rural)
Low	**AURA** ↓	0	8
	GUIDED MARL	4	9
	MARL-ONLY	11	34
Normal	**AURA**↓	29	83
	GUIDED MARL	84	98
	MARL-ONLY	122	209
High	**AURA**↓	472	381
	GUIDED MARL	514	405
	MARL-ONLY	846	612

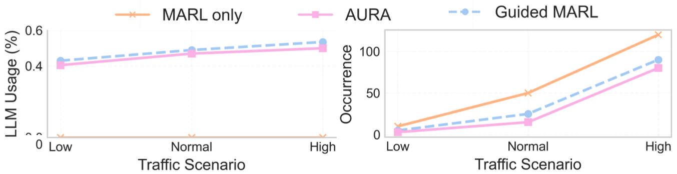
(a) Dropped requests across traffic scenarios.

Traffic Condition	Agent 1	Agent 2
Normal	0.0135*	0.0018**
High	0.0032**	0.0018**

(b) p-values (Kruskal-Wallis test)

	MARL-only vs Guided		MARL-only vs AURA	
	Agent 1 (urban)	**Agent 2 (rural)**	**Agent 1 (urban)**	**Agent 2 (rural)**
Low	0.370	0.090	0.170	0.086
Normal	0.033*	0.013*	0.005**	0.001***
High	0.004**	0.019*	0.003**	0.001***

(c) Post-hoc p-values (Dunn's test with Holm correction)



(d) Agent reliance on LLM suggestions.

(e) System failure.

Figure 2: Comparison of MARL-ONLY, GUIDED MARL, and AURA across traffic conditions. (a) Dropped requests, with largest gains under normal and high traffic. (b–c) Statistical analysis showing strongest improvements for AURA and GUIDED MARL (* $p < .05$, ** $p < .01$, *** $p < .001$). (d) LLM usage rates, indicating moderate reliance. (e) System failure counts, reduced by both LLM-guided methods under higher traffic.

References

Ammar Kamal Abasi, Moayad Aloqaily, Mohsen Guizani, and Bassem Ouni. Metaheuristic algorithms for 6G wireless communications: Recent advances and applications. *Ad Hoc Networks*, page 103474, 2024.

Soheil Abbasloo, Chen-Yu Yen, and H Jonathan Chao. Classic meets modern: A pragmatic learning-based congestion control for the internet. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, pages 632–647, 2020.

Hwijeen Ahn, Seongjin Shin, Sang-Woo Lee, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung Park, Jung-Woo Ha, and Nako Sung. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5166–5180. Association for Computational Linguistics, 2022.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Mohammed Banafaa, Ibraheem Shayea, Jafri Din, Marwan Hadri Azmi, Abdulaziz Alashbi, Yousef Ibrahim Daradkeh, and Abdulraqeb Alhammadi. 6G mobile communication technology: Requirements, targets, applications, challenges, advantages, and opportunities. *Alexandria Engineering Journal*, 64:245–274, 2023.

Abdelhak Bentaleb, Christian Timmerer, Ali C Begen, and Roger Zimmermann. Bandwidth prediction in low-latency chunked streaming. In *Proceedings of the 29th ACM workshop on network and operating systems support for digital audio and video*, pages 7–13, 2019.

Rupali Bhati, Sai Krishna Gottipati, Clodéric Mars, and Matthew E Taylor. Curriculum learning for cooperation in multi-agent reinforcement learning. *arXiv preprint arXiv:2312.11768*, 2023.

Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015.

Daniela M Casas-Velasco, Oscar Mauricio Caicedo Rendon, and Nelson LS da Fonseca. Intelligent routing based on reinforcement learning for software-defined networking. *IEEE Transactions on Network and Service Management*, 18(1):870–881, 2020.

Robin Chataut, Mary Nankya, and Robert Akl. 6G networks and the AI revolution—exploring technologies, applications, and emerging challenges. *Sensors*, 24(6):1888, 2024.

Li Chen, Justinas Lingys, Kai Chen, and Feng Liu. Auto: Scaling deep reinforcement learning for datacenter-scale automatic traffic optimization. In *Proceedings of the 2018 conference of the ACM special interest group on data communication*, pages 191–205, 2018.

Tianshu Chu, Jie Wang, Lara Codecà, and Zhaojian Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE transactions on intelligent transportation systems*, 21(3):1086–1095, 2019.

Qimei Cui, Xiaohu You, Ni Wei, Guoshun Nan, Xuefei Zhang, Jianhua Zhang, Xinchen Lyu, Ming Ai, Xiaofeng Tao, Zhiyong Feng, et al. Overview of AI and communication for 6G network: Fundamentals, challenges, and future research opportunities. *arXiv preprint arXiv:2412.14538*, 2024.

Alexis Dinno. Nonparametric pairwise multiple comparisons in independent groups using dunn's test. *The Stata Journal*, 15(1):292–300, 2015.

Anutusha Dogra, Rakesh Kumar Jha, and Kumud Ranjan Jha. Intelligent routing for enabling haptic communication in 6g network. In *2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 419–422. IEEE, 2023.

Amal Feriani and Ekram Hossain. Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial. *IEEE Communications Surveys & Tutorials*, 23(2):1226–1252, 2021.

Matthew J Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable MDPs. In *AAAI fall symposia*, volume 45, page 141, 2015.

Long He, Geng Sun, Dusit Niyato, Hongyang Du, Fang Mei, Jiawen Kang, Mérouane Debbah, and Zhu Han. Generative AI for game theory-based mobile networking. *IEEE Wireless Communications*, 32(1):122–130, 2025.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022.

Muhammad Jamshaid Iqbal, Muhammad Farhan, Farhan Ullah, Gautam Srivastava, and Sohail Jabbar. Intelligent multimedia content delivery in 5G/6G networks: a reinforcement learning approach. *Transactions on Emerging Telecommunications Technologies*, 35(4):e4842, 2024.

Fahime Khoramnejad and Ekram Hossain. Generative AI for the optimization of next-generation wireless networks: Basics, state-of-the-art, and open challenges. *IEEE Communications Surveys & Tutorials*, 2025.

Wonjun Kim, Yongjun Ahn, Jinhong Kim, and Byonghyo Shim. Towards deep learning-aided wireless channel estimation and channel state information feedback for 6G. *Journal of Communications and Networks*, 25(1): 61–75, 2023.

William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Forty-first International Conference on Machine Learning*.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. A survey of reinforcement learning informed by natural language. *arXiv preprint arXiv:1906.03926*, 2019.

Madduma Wellalage Pasan Maduranga, Valmik Tilwari, RMMR Rathnayake, and Chamali Sandamini. Ai-enabled 6g internet of things: Opportunities, key technologies, challenges, and future directions. In *Telecom*, volume 5, pages 804–822. MDPI, 2024.

Ziadoon K Maseer, Qusay Kanaan Kadhim, Baidaa Al-Bander, Robiah Yusof, and Abdu Saif. Meta-analysis and systematic review for anomaly network intrusion detection systems: Detection methods, dataset, validation methodology, and challenges. *IET Networks*, 2024.

Zepeng Ning and Lihua Xie. A survey on multi-agent reinforcement learning and its application. *Journal of Automation and Intelligence*, 2024.

Hafiz Muhammad Fahad Noman, Effariza Hanafi, Kamarul Ariffin Noordin, Kaharudin Dimyati, Mhd Nour Hindia, Atef Abdrabou, and Faizan Qamar. Machine learning empowered emerging wireless networks in 6g: Recent advancements, challenges & future trends. *IEEE Access*, 2023.

Narjes Nourzad and Bhaskar Krishnamachari. Smart routing with precise link estimation: DSEE-based anypath routing for reliable wireless networking. *arXiv preprint arXiv:2405.10377*, 2024.

OpenAI. Chatgpt. `https://chat.openai.com/chat`, 2024.

Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys (CSUR)*, 54(4):1–34, 2021.

Md Shahjalal, Woojun Kim, Waqas Khalid, Seokjae Moon, Murad Khan, ShuZhi Liu, Suhyeon Lim, Eunjin Kim, Deok-Won Yun, Joohyun Lee, et al. Enabling technologies for ai empowered 6g massive radio access networks. *ICT Express*, 9(3):341–355, 2023.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Neural Information Processing Systems*, 2023. URL `https://api.semanticscholar.org/CorpusID:258833055`.

Chuanneng Sun, Songjun Huang, and Dario Pompili. Hmaac: Hierarchical multi-agent actor-critic for aerial search with explicit coordination modeling. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7728–7734. IEEE, 2023.

Chuanneng Sun, Songjun Huang, and Dario Pompili. LLM-based multi-agent reinforcement learning: Current and future directions. *arXiv preprint arXiv:2405.11106*, 2024.

Karl Tuyls. Multiagent learning: From fundamentals to foundation models. AAMAS '23, page 1, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.

Karl Tuyls and Gerhard Weiss. Multiagent learning: Basics, challenges, and prospects. *Ai Magazine*, 33(3): 41–41, 2012.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Caroline Wang, Ishan Durugkar, Elad Liebman, and Peter Stone. Dm^2: Decentralized multi-agent reinforcement learning via distribution matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11699–11707, 2023a.

Jiadai Wang, Jiajia Liu, Jingyi Li, and Nei Kato. Artificial intelligence-assisted network slicing: Network assurance and service provisioning in 6g. *IEEE Vehicular Technology Magazine*, 18(1):49–58, 2023b.

Xiangwen Wang, Xianghong Lin, and Xiaochao Dang. Supervised learning in spiking neural networks: A review of algorithms and evaluations. *Neural Networks*, 125:258–280, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Duo Wu, Panlong Wu, Miao Zhang, and Fangxin Wang. Mansy: Generalizing neural adaptive immersive video streaming with ensemble and representation learning. *IEEE Transactions on Mobile Computing*, 2024.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Zijiang Yan, Hao Zhou, Hina Tabassum, and Xue Liu. Hybrid llm-ddqn based joint optimization of v2i communication and autonomous driving. *IEEE Wireless Communications Letters*, 2025.

Hanqing Yang, Jingdi Chen, Marie Siew, Tania Lorido-Botran, and Carlee Joe-Wong. Llm-powered decentralized generative agents with adaptive hierarchical knowledge graph for cooperative planning. *arXiv preprint arXiv:2502.05453*, 2025.

Helin Yang, Arokiaswami Alphones, Zehui Xiong, Dusit Niyato, Jun Zhao, and Kaishun Wu. Artificial-intelligence-enabled intelligent 6g networks. *IEEE network*, 34(6):272–280, 2020.

Yi Yang, Fenglei Li, Xinzhe Zhang, Zhixin Liu, and Kit Yan Chan. Dynamic power allocation in cellular network based on multi-agent double deep reinforcement learning. *Computer Networks*, 217:109342, 2022.

Alessio Zappone, Marco Di Renzo, and Mérouane Debbah. Wireless networks design in the era of deep learning: Model-based, ai-based, or both? *IEEE Transactions on Communications*, 67(10):7331–7376, 2019.

Yuan Zhuang, Yi Shen, Zhili Zhang, Yuxiao Chen, and Fei Miao. YOLO-MARL: You only llm once for multi-agent reinforcement learning. *arXiv preprint arXiv:2410.03997*, 2024.

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.

Hang Zou, Qiyang Zhao, Lina Bariah, Yu Tian, Mehdi Bennis, Samson Lasaulce, Merouane Debbah, and Faouzi Bader. GenAINet: Enabling wireless collective intelligence via knowledge transfer and reasoning. *arXiv preprint arXiv:2402.16631*, 2024.

A Related Work

Artificial intelligence (AI) has been extensively used in cellular networks to address challenges in network optimization. In particular, supervised and unsupervised learning techniques, especially deep learning, have shown considerable potential to improve various aspects of network operations [Zappone et al., 2019, Maseer et al., 2024]. Supervised learning has been widely utilized when it comes to predictive networking tasks, including traffic classification and bandwidth prediction. By training deep neural networks (DNNs) on historical data, these methods enable automated decision-making to optimize system performance [Wang et al., 2020, Bentaleb et al., 2019]. Nonetheless, the growing complexity and evolving demands of 6G networks expose fundamental limitations in these approaches. Common issues include scalability and limited adaptability to real-time fluctuations. Moreover, the diversity of networking tasks further limits model reuse, as each problem often requires a distinct architecture [Ren et al., 2021]. Recent advances, such as structured Transformers [Vaswani, 2017], have improved adaptability. Even so, they still rely on manual tuning and architectural modifications [Wu et al., 2024], increasing the complexity of the deployment and engineering costs.

In response to the challenges of conventional learning approaches, researchers have increasingly turned to reinforcement learning (RL) to meet the dynamic requirements of modern networks [Chen et al., 2018]. RL has been applied to various areas spanning congestion control, traffic optimization, and cloud cluster job scheduling (CJS) [Abbasloo et al., 2020, Casas-Velasco et al., 2020, Nourzad and Krishnamachari, 2024]. Multi-agent RL (MARL) extends these capabilities by enabling collaboration among multiple agents, offering improved scalability and robustness in distributed systems [Yang et al., 2022, Ning and Xie, 2024, Bloembergen et al., 2015]. However, while these techniques perform well in controlled settings, real-world deployments present significant challenges. Computational inefficiencies slow down processing, high-dimensional environments hinder convergence, and dynamic conditions with unpredictable user behaviors make adaptation difficult [Iqbal et al., 2024, Tuyls and Weiss, 2012].

Recent research has begun exploring the integration of MARL with Large Language Models (LLMs) to improve coordination and decision-making [Sun et al., 2024, Tuyls, 2023]. Leveraging the advanced planning and reasoning capabilities of LLMs, researchers have demonstrated that these models can coordinate multiple agents within a network [Zou et al., 2024, Yan et al., 2025, He et al., 2025]. On top of that, LLMs can uncover patterns in large datasets that traditional RL methods may overlook. These capabilities simplify collective decision-making and enable real-time adaptation to fluctuating user demands in dynamic network scenarios [Huang et al., 2022]. Furthermore, by integrating multimodal data and anticipating future network states, LLMs can optimize resource allocation to enhance both operational efficiency and network robustness in 6G environments [Luketina et al., 2019, Zou et al., 2023, Khoramnejad and Hossain, 2025]. This hybrid approach effectively bridges critical gaps in adaptability and decision-making under dynamic conditions. Regardless, its practical application remains under-explored, particularly in the context of multi-agent collaboration and real-time resource optimization, where scalability and efficiency are key.

B Reward Algorithm

Algorithm 1 describes MARL training under LLM guidance, where agents combine local policies with LLM-suggested actions using a trust score. Immediate rewards update Q-values in real time, while accumulated histories enable periodic delayed rewards that reinforce alignment with high-level objectives.

C Experimental Setup

C.1 MARL formulation.

We model the system as a Partially Observable Markov Decision Process (POMDP), which more accurately reflects the environment compared to a fully observable MDP [Hausknecht and Stone, 2015]. The LLM has a global, though partial, understanding of the network's state based on various data inputs (e.g., social media and network statistics). While the LLM has access to a broader set of information in comparison to individual agents, it still deals with partial observability since it doesn't have complete information about the environment (like the exact state of every user or base station). Similarly, the MARL agents operate with partial observability, since each agent is limited to their local observations (e.g., signal strength, traffic load). The agents must make decisions with incomplete knowledge of the broader network. In this setup, LLM serves as a centralized decision-maker that breaks down high-level goals into subgoals for the agents to execute locally. In other words, the LLM centrally reasons and provides strategic guidance, making it more aligned with a centralized POMDP, where the central entity (LLM) orchestrates the strategy based on partial information. We now turn to defining the state, action, and reward spaces for the MARL agents within the AURA system:

- OBSERVATION SPACE: Each agent's state space is defined by a set of local features that capture network conditions relevant to decision-making: (1) the transmission *power level* of the agent's cell tower, (2) the quality

Algorithm 1 MARL Training with LLM Guidance

```
function DELAYED_REWARD($\mathcal{A}$)
    for each $agent_i \in \mathcal{A}$ do
        $r_{delayed,i} \leftarrow$ COMPUTE_DELAYED_REWARD($\mathcal{H}_i$)
        for each $(s, a) \in \mathcal{H}_i$ do    ▷ Apply rewards uniformly to all past experiences in history
            $Q(s, a) \leftarrow Q(s, a) + \alpha \cdot r_{delayed,i}$    ▷ $\alpha$ controls how much the delayed reward influences learning
        end for
        CLEAR_HISTORY($\mathcal{H}_i$)
    end for
end function
for $epoch$ in range($N$) do
    for each $agent_i \in \mathcal{A}$ do
        $a_{i,LLM} \leftarrow \pi$.ACTION($s_i, s_{-i}$)    ▷ Action recommended to $agent_i$ by LLM $\pi$
        $a_{i,RL} \leftarrow agent_i$.ACTION($s_i$)    ▷ Action taken by $agent_i$ based on local, real-time information
        $a_i \leftarrow agent_i$.COMBINE_DECISION($a_{i,LLM}, a_{i,RL}$, trustscore$_i$)
        $s'_i, r_{immediate,i}, \psi_i \leftarrow$ TAKE_ACTION($a_i$)    ▷ $\psi_i$ captures environmental parameters like SNR, signal strength, etc.
        $Q(s_i, a_i) \leftarrow Q(s_i, a_i) + \alpha \left[r_{immediate,i} + \gamma \max_{a'} Q(s'_i, a') - Q(s_i, a_i)\right]$    ▷ Q-learning update using immediate rewards
        $agent_i$.UPDATE_TRUST($\pi, s_i, a_{i,LLM}, a_{i,RL}, r_{immediate,i}, a_i$)
    end for
end for
```

of *network coverage* (categorized as good, fair, or poor), (3) the user *connection status*, which indicates whether the tower has reached its maximum capacity, and (4) number of calls dropped from last time. Based on these observations, agents must select actions that optimize network performance while adapting to environmental changes.

- ACTION SPACE: The available action space consists of four primary decisions. Agents can (1) *increase* transmission power when signal strength is weak and capacity allows. They can (2) *decrease* power to reduce interference or conserve energy. If network conditions are stable, they may (3) *maintain* the current state. Finally, they can initiate a (4) *handoff* request when a user's connection quality deteriorates, provided a neighboring tower offers a better alternative.

- REWARD: The reward function combines both *immediate* and *delayed* rewards. Immediate rewards are assigned at each time step, providing feedback based on local performance indicators such as connectivity quality and energy efficiency. Delayed rewards, on the other hand, are computed every few episodes by the CAC using a language model-based evaluation mechanism. The delayed reward function processes the agent's historical state-action trajectory to derive a performance-based reward score, considering factors such as alignment with high-level objectives and overall network impact. This additional feedback mechanism helps agents optimize long-term behavior rather than overfitting to short-term gains.

C.2 LLM Query Design.

- CENTRALIZED LANGUAGE MODEL: In our implementation, a detailed prompt is constructed to guide the LLM in suggesting actions based on the network condition.. The prompt is composed by incorporating key metrics from both the neighboring and target cell towers, including power levels, coverage quality, and user connection status. We use *Claude-sonnet-4* as the underlying model, which generates responses based on these inputs. Then, a translator module extracts a clean numeric code from the LLM's response, ensuring that only the intended output is used by the MARL framework. This approach allows the LLM to focus solely on processing the relevant state information and providing concise, actionable feedback, while the broader action execution is managed by the MARL framework.

- CENTRALIZED ALIGNMENT CONTROLLER: As described earlier, the Centralized Alignment Controller (CAC) plays two roles in our framework. However, at this stage of our work, we focus on its function as a reward provider, evaluating MARL agents' performance and assigning rewards accordingly. To implement this, we use an LLM-based evaluation mechanism that instructs the model, to act as an expert evaluator. The evaluation considers factors such as network efficiency, fairness, adaptability, and long-term performance. Based on these criteria, a reward is assigned within the range of $[-1, 1]$. A score of +1 indicates "*Excellent*" optimization (i.e., maximized efficiency, user experience, and adaptability), while -1 reflects "*Poor*" optimization (i.e., significant issues like congestion, dropped connections, or poor resource use). Intermediate values represent varying levels of improvement or decline.

Prompts and implementation details can be found in: https://anonymous.4open.science/r/AURA-F79F/.