Against the Monolithic Wireless World Model: Why NextG Needs Composable and Agentic Intelligence

Aladin Djuhera [1] **Farhan Ahmed** [2] **Vlad Andrei** [1] **Swanand Ravindra Kadhe** [2] **Alecio Binotto** [3] **Haris Gacanin** [4] **Holger Boche** [1]

Abstract

AI-native 6G visions increasingly invoke wireless foundation models, large multimodal models, and wireless world models as the natural endpoint of AI-native networking, drawing an analogy to recent developments in large language models (LLMs). We argue that this analogy is structurally incomplete. The success of LLMs is based on a broad, reusable, and largely self-contained tokenized data substrate, whereas the wireless domain lacks an equivalent data foundation. Unlike text, code, or images, wireless data such as CSI tensors, IQ samples, or scheduler logs are not self-contained: their meaning is configuration-dependent, simulator-conditioned, task-disaggregated, and weakly grounded in operational feedback, all structural bottlenecks that undermine current pre- and post-training recipes. We therefore argue that monolithic models, including mixture-of-experts (MoE) and wireless world models, are not the most realistic near-term path toward deployable AI-native networks. Instead, emerging evidence points toward composable and agentic network architectures, where general reasoning models orchestrate specialized signal processing models, classical algorithms, digital twins, standards-aware retrieval, and safety checks through explicit programmable interfaces.

1. Introduction: Hype and Reality

AI-native visions of 6G and beyond describe a future in which learning systems permeate every layer of the wireless protocol stack, from physical-layer signal processing to network management and application-level semantics (Saad et al., 2025; Jiang et al., 2025). This vision has catalyzed serious research into foundation models for several wireless applications, including channel estimation, task prediction, and sensing (Djuhera et al., 2026a; Liu et al., 2025a; Sheng et al., 2025; Yang et al., 2025). Continuing this wave, many works are proposing so-called large wireless models (LWMs) (Alikhani et al., 2025), large multimodal models (LMMs) (Xu et al., 2024), and, more recently, wireless world models (WWMs) (Zou et al., 2026b), motivated by novel architectures like JEPA (Assran et al., 2023). Such proposals advocate for training models that embed *universal wireless intelligence* across all layers, similar as to how state-of-the-art LLMs seem to exhibit beyond-human-level intelligence across several diverse tasks.

> *But many years later, where is the ChatGPT equivalent to wireless foundation model intelligence?*

We argue that the LLM analogy is structurally incomplete and incompatible with the wireless protocol stack. In particular, LLMs and world models alike work not only because of sheer model scale but, more importantly, because of broad, reusable, and self-contained tokenized data (Kaplan et al., 2020). However, wireless AI has neither ingredient in adequate form: data samples are almost always fragmented across tasks, configurations, simulators, protocol layers, and operational contexts, distinguishing them sharply from text and code. This is *not* because the community has been insufficiently ambitious, but because of the high degree of modularity when designing wireless networks at scale.

In this paper, we make three main claims:

1. **The current AI-native direction is under-specified**: many works quickly adopt new AI architectures without defining compatible data interfaces, operational boundaries, or paths through existing telecom infrastructure.

2. **Wireless AI data suffers from four main bottlenecks**: configuration dependence, simulation and setup dependence, no universal wireless token, and missing operational feedback are detrimental to model development.

[1]Technical University Munich [2]IBM Research [3]Zeiss AG [4]RWTH Aachen University. Correspondence to: Aladin Djuhera <aladin.djuhera@tum.de>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

3. **AI-native wireless must be agentic and compositional**: the protocol stack should not be replaced by monolithic all-purpose models, but instead, should combine agentic reasoning and tool calling with specialized models.

We are not claiming wireless foundation models are useless or that LLMs have no role in wireless systems. Our claim is narrower and constructive: the near-term path to AI-native NextG should be *compositional rather than monolithic*, with *agent harnesses* that orchestrate configuration-aware, provenance-annotated wireless data, specialized tools, and models through explicit interfaces.

2. AI-Native NextG is Underspecified

The first wave of wireless AI produced valuable task-specific models. Neural networks have been successfully trained for channel estimation (Jin et al., 2019), beam prediction (Li et al., 2023), and resource allocation (Djigal et al., 2022), showing that wireless intelligence can be embedded in specific tasks and control loops. This is also reflected in the recent 3GPP Release-18/5G-Advanced work on AI/ML for the NR air interface (Lin et al., 2023), which emphasizes concrete use cases with implementation details and, most importantly, clear interface specifications for AI modules.

The current wave of AI aims to create foundation models that provide intelligence beyond the per-task level. This implies models that generalize across configurations, protocol settings, and simulation environments. However, the operational meaning of such models is often vague or left underspecified. *What is the wireless analogue of a tokenized corpus? Which observations, configurations, actions, and feedback signals should be included? At which layer of the protocol stack should the model operate?* Without answering these questions, "AI-native" risks becoming a label for applying generic post-training to telecom data, rather than a concrete architecture for deployable network intelligence.

Early telecom LLM efforts illustrate this limitation. Fine-tuning models on standards documents and technical reports can improve domain question answering (Q&A) and standards retrieval (Maatouk et al., 2024; Nikbakht et al., 2024), but they remain primarily language-level systems and thus cannot by themselves learn the dynamics of wireless channels or solve complex mobility or scheduling problems. Furthermore, it still remains unclear how to measure actionable wireless intelligence due to the lack of adequate benchmarks. Only recently have the GSMA Open-Telco LLM Benchmarks (GSMA, 2025) included operational tasks such as configuration generation, troubleshooting, and quantitative reasoning. Yet they remain mostly language-level evaluations, while benchmarks for wireless world models are still missing. Even in this limited setting, strong frontier models struggle with schema-compliant intent-to-configuration, supporting GSMA's conclusion that agentic architectures are needed. This aligns with our position: deployable telecom AI requires explicit interfaces between foundation models and specialized solvers, as well as meaningful benchmarks. We provide a broader discussion of related work in App. A.

3. Structural Bottlenecks for Wireless AI Data

Current approaches to wireless LLMs and world models are difficult to reconcile with established training methods because they face structural data limitations. Wireless data is often underspecified, highly context-dependent, and difficult to universally tokenize, making it unclear whether standard scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) transfer to this domain. As a result, both LLMs and world models risk offering limited deployable value unless data interfaces, operational scope, and feedback mechanisms are explicitly defined. We identify four main bottlenecks:

Bottleneck 1: Configuration Dependence. In wireless settings, a data sample is inseparable from the system configuration that produced it. The same learning problem changes entirely under different carrier frequencies, antenna geometries, beam codebooks, pilot designs, numerologies, and schedulers. Thus, a CSI tensor labeled *"good channel state"* under one configuration is meaningless under another, and training across all plausible configurations is combinatorially infeasible. This leaves wireless AI models with three unpalatable options: (i) *infer the configuration from the signal*, which works in-distribution but fails silently outside it, (ii) *condition explicitly on configuration metadata*, which is rarely done systematically or even available, or (iii) *average over a mix of configurations*, yielding models that are mediocre everywhere and reliable nowhere. 3GPP's focus on concrete air-interface use cases reflects precisely this constraint and aims to isolate tasks to make their data regime tractable. Thus, a wireless dataset without configuration metadata is a dataset for interpolation, not intelligence.

Bottleneck 2: Simulation and Setup Dependence. Wireless AI relies heavily on simulation because real deployment data is scarce, proprietary, and difficult to label. Platforms such as DeepMIMO (Alkhateeb, 2019), Sionna (Hoydis et al., 2022), OpenRAN Gym (Bonati et al., 2023), and Colosseum (Polese et al., 2024) provide essential infrastructure for data generation, but every simulator also encodes its own assumptions about signal propagation, interference, and protocol abstractions. Thus, models trained on one simulator may learn its artifacts rather than the underlying wireless physics, echoing the sim-to-real problem in robotics. Thus, unlike LLMs, they cannot be easily downloaded and reused across arbitrary configurations, simulators, and deployments. Wireless AI therefore requires explicit setup provenance, including the simulator name and version, scenario, channel

model, configuration, impairments, and random seeds.

Bottleneck 3: No Universal Wireless Token. LLMs and world models are successful in encoding vast knowledge in part because text tokens are broadly reusable abstractions. However, IQ samples, CSI tensors, beam indices, and scheduler traces differ not only in format, but also in abstraction level, control timescale, and causal regime. *A CSI tensor is not a sentence. An IQ stream is not a paragraph. A scheduler trace is not source code.* Moreover, the task context often lives outside the sample, e.g., a CSI tensor alone says little about antenna geometry, pilot structure, feedback protocol, or mobility model. Thus, before asking how to train a wireless foundation model, one must answer: *What is the wireless equivalent of a token, instruction, and tool call?* Although recent works attempt to tokenize or unify wireless streams (Liu et al., 2024; Djuhera et al., 2026a), these representations remain task- and modality-specific. This heterogeneity is significant: PHY receivers, MAC schedulers, and RAN optimizers operate at different timescales, with different losses, action spaces, latency constraints, and safety requirements. The wireless stack is modular for a reason and AI should respect those boundaries unless deployment evidence justifies breaking them.

Bottleneck 4: Missing Operational Feedback. Modern LLMs improve through instruction tuning, retrieval feedback, and preference optimization (Ouyang et al., 2022; Lewis et al., 2020). However, post-training alignment as understood for LLMs has no wireless counterpart, because there is no scalable notion of operator or user preference over PHY/MAC/RAN actions, and no safe way to learn one from live trial and error. Operators cannot freely A/B test schedulers on live users, standards and vendor constraints limit admissible actions, and the most informative failures are precisely those too costly to induce. World models do not resolve this either and only relocate the problem, since the model itself must be calibrated against operational ground truth data that the same missing feedback channel is supposed to provide. Thus, our position remains that the only architecture compatible with the operational data output of wireless networks is an agentic one, in which reasoning models consume structured data (telemetry, KPIs, configuration logs) through explicit interfaces rather than learning from a preference signal that does not exist.

Given these bottlenecks, it is evident that wireless datasets should be accompanied by data cards that document configuration, provenance, and other assumptions (see App. B).

4. The Case Against Wireless World Models

A wireless world model is a learned simulator of radio network dynamics (channels, queues, KPIs), conditioned on operator actions such as beam updates, handover policies, or resource allocations (see App. C). It is intended to support counterfactual rollouts (e.g., *"What if we changed the scheduler?"*) without interacting with the live network. This makes world models attractive for sample-efficient and proactive planning that classical reinforcement learning (RL) cannot offer. However, we argue that the structural data bottlenecks are inherited in world models and that they are compelling only when the "world" is explicitly scoped. Indeed, recent proposals already point toward an agentic composition rather than monolithic pretraining (Zou et al., 2026b). We make the following claims:

The Wireless World is Not One World. In the wireless domain, the relevant "world" changes with the control problem. For beam tracking, it may consist of channel states, mobility, blockage, and beam codebooks over millisecond horizons. For network slicing, it may consist of queues, PRB allocations, transport congestion, and SLA margins over seconds or minutes. These are not merely different features of one homogeneous state space, but correspond to different abstraction levels, timescales, action spaces, and objective functions. A single world model that tries to absorb all of them must either flatten these distinctions into one opaque latent state or hand-design interfaces between them. The former makes the model difficult to interpret and validate, while the latter already turns the system into a modular architecture. Thus, the problem is not that wireless worlds cannot be defined, but that useful definitions are tied to local predictive models for particular control loops, defying the notion of a single wireless world model. Moreover, such models inherit rather than resolve the data bottlenecks: action-conditioned rollouts require logged trajectories, spatial prediction requires configuration-aware RF data, synthetic pretraining requires simulator provenance, and cross-layer modeling still lacks a shared tokenization. World models therefore make all four bottlenecks load-bearing at once.

World Models Become Agentic When Made Practical. World models may be valuable for traffic generation, digital-twin acceleration, and counterfactual planning (Zheng et al., 2026), but they should not be treated as default replacements for robust channel estimators, schedulers, or operator workflows. Recent work (Zou et al., 2026b) acknowledges this and proposes to decompose wireless world models into three layers: (i) *a Field World Model layer* for spatial prediction, (ii) *a Control World Model layer* for action-conditioned KPI rollouts, and (iii) *a Foundation Model layer* for intent translation and orchestration. The latter is explicitly agentic: it coordinates calls to world-model components, simulators, digital twins, O-RAN xApps/rApps, policy engines, and constraint solvers, while validating candidate actions against SLA and safety constraints before execution. Thus, the most developed case for wireless world models is

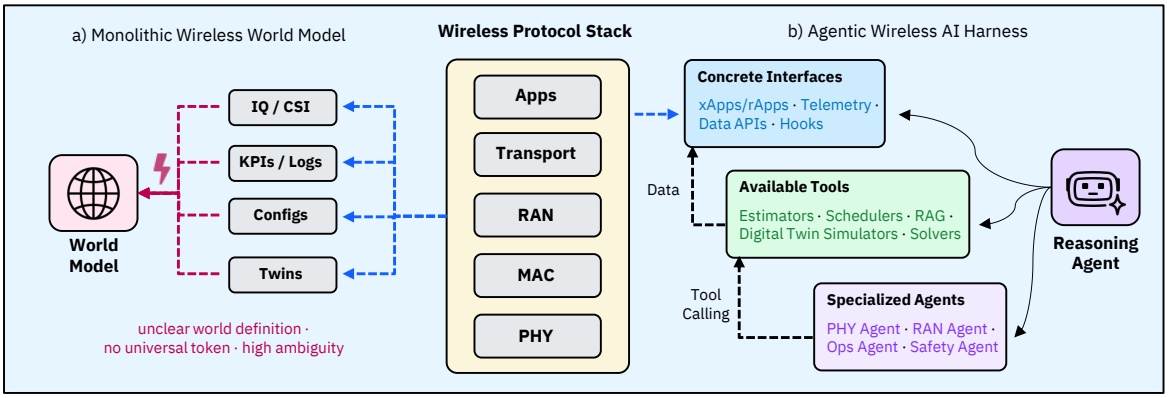

Figure 1. Monolithic wireless world models collapse heterogeneous data into an ambiguous interface, whereas agent harnesses orchestrate tools, interfaces, and specialized agents through explicit and auditable tool calls. See App. D for a detailed architecture description.

not one for monolithic intelligence, but for compositional, tool-using architectures compatible with existing telecom infrastructure, thereby undermining initial definitions of a single, unitary wireless world model.

The strongest argument *for* wireless world models therefore becomes an argument against treating them as the endpoint: they motivate an agentic architecture in which world-model components are tools, not the system itself.

5. A Composable and Agentic Alternative

If monolithic models are the wrong abstraction, the alternative is not to abandon them, but to place them inside an *agent harness*, a runtime layer in which reasoning agents orchestrate specialized wireless models, classical algorithms, simulators, telemetry databases, retrieval systems, and safety monitors through explicit interfaces, rather than replacing the protocol stack (see Fig. 1). This architecture is more aligned with how wireless systems are already built.

Specifically, PHY, MAC, RAN, transport, core, and management layers are highly modular and O-RAN further exposes this modularity through near-real-time and non-real-time control loops, xApps/rApps, telemetry streams, and policy interfaces (Bonati et al., 2023), providing the necessary connectors to agentic workflows (Dev et al., 2025). Thus, PHY tasks can remain with specialized signal-processing models, RAN optimization can be handled by scoped xApps/rApps, digital twins can provide counterfactual validation, and standards retrieval can constrain configuration logic. Within the harness, LLM-based reasoning agents are equipped with tool-use capabilities (Masterman et al., 2024) and translate intents, plan multi-step workflows, select tools, and escalate uncertain cases to human operators. The agentic layer thus enables guardrailed orchestration that monolithic world models cannot without becoming modular themselves.

Furthermore, the agentic approach also directly addresses the wireless data bottlenecks in Sec. 3. First, configuration dependence is handled by making configuration an explicit input to each tool call rather than an implicit variable hidden inside a latent model. Second, simulation dependence is made auditable because simulator identity, scenario, channel model, and random seed can be attached to every generated rollout. Third, the lack of a universal wireless token is avoided: CSI tensors, IQ streams, KPI logs, topology graphs, and standards documents need not be forced into one representation, but can remain in their native formats behind specialized interfaces. Fourth, missing operational feedback is mitigated by separating recommendation from execution. The corresponding agent harness can run shadow-mode analyses, compare actions in digital twins, enforce guardrails, and request operator approval before affecting live users. In addition, new agent skills can be added modularly as use cases evolve and as future 3GPP releases introduce new interfaces, procedures, and requirements.

For example, consider a cell experiencing degraded beam-management performance under high mobility. An agentic system first retrieves the cell configuration, beam codebook, UE mobility traces, and recent KPI degradation. It then calls a classical beam-selection baseline and a learned beam predictor, compares their confidence and latency, and tests both under nearby mobility scenarios in a digital twin. If the learned predictor is confident and passes the simulator and safety checks, the agent can recommend switching the corresponding xApp to the learned model. Thus, wireless modules are not replaced globally or permanently, but selected dynamically according to the environment. Further examples and detailed architectures are provided in App. D.

Consequently, intelligence arises from composition rather than from forcing one model to internalize the entire wireless stack. This architecture also remains adaptable because individual models, tools, and workflows can be replaced, reconfigured, or validated independently as network conditions, standards, and deployment requirements evolve.

6. Conclusion

In this paper, we argued that monolithic wireless world models face structural data bottlenecks that scale alone cannot resolve, thereby limiting their broad usability. As an alternative, we advocate composable AI-native wireless systems, in which agent harnesses coordinate specialized wireless models and classical solvers through explicit interfaces. Our position is thus clear: AI-native wireless should not be measured by whether the community can train ever larger models, but whether wireless intelligence can be made deployable, configuration-aware, and, most importantly, compatible with the modular protocol stack.

Acknowledgments

This work was supported in part by the German Federal Ministry of Research, Technology and Space (BMFTR) within the research hubs 6G-life (Grant 16KISK002) and 6GEM+ (Grant 16KIS2409K) through the projects GEM-X (Grant 16KISS004K), VICTOR6G (Grant 16KIS2547), QD-CamNetz (Grant 16KISQ077), and AISAC (Grant 16KIS2462). This work was also supported by the Bavarian Ministry of Science and the Arts through the project Next Generation AI Computing (gAIn), by the Bavarian Ministry of Economic Affairs, Regional Development and Energy through the project 6G Future Lab Bavaria, by the German Research Foundation (DFG) within the Centre for Tactile Internet with Human-in-the-Loop (CeTI) as part of Germany's Cluster of Excellence (EXC 2050/2, ID 390696704), and in part by IBM Research.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning and wireless communications. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Alikhani, S., Charan, G., and Alkhateeb, A. Large Wireless Model (LWM): A Foundation Model for Wireless Channels. *arXiv preprint arXiv:2411.08872*, 2024.

Alikhani, S., Charan, G., and Alkhateeb, A. Large Wireless Model (LWM): A Foundation Model for Wireless Channels, 2025. URL https://arxiv.org/abs/2411.08872.

Alkhateeb, A. DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications. *arXiv preprint arXiv:1902.06435*, 2019.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Bonati, L., Polese, M., D'Oro, S., Basagni, S., and Melodia, T. OpenRAN Gym: AI/ML Development, Data Collection, and Testing for O-RAN on PAWR Platforms. *Computer Networks*, 220:109502, 2023.

Cheng, C., Guo, L., Wu, T., Sun, J., Gui, G., Adebisi, B., Gacanin, H., and Sari, H. Machine-Learning-Aided Trajectory Prediction and Conflict Detection for Internet of Aerial Vehicles. *IEEE Internet of Things Journal*, 9(8): 5882–5894, 2021.

Dev, K., Khowaja, S. A., Zeydan, E., Singh, K., and Debbah, M. Advanced Architectures Integrated With Agentic AI for Next-Generation Wireless Networks. *IEEE Communications Standards Magazine*, 2025.

Djigal, H., Xu, J., Liu, L., and Zhang, Y. Machine and Deep Learning for Resource Allocation in Multi-Access Edge Computing: A Survey. *IEEE Communications Surveys & Tutorials*, 24(4):2449–2494, 2022.

Djuhera, A., Andrei, V. C., Li, X., Mönich, U. J., Boche, H., and Saad, W. R-SFLLM: Jamming Resilient Framework for Split Federated Learning with Large Language Models. *IEEE Transactions on Information Forensics and Security*, 2025a.

Djuhera, A., Andrei, V. C., Pourghasemian, M., Gacanin, H., Boche, H., and Saad, W. R-MTLLMF: Resilient Multi-Task Large Language Model Fusion At The Wireless Edge. In *ICC 2025-IEEE International Conference on Communications*, pp. 1554–1559. IEEE, 2025b.

Djuhera, A., Kadhe, S. R., Ahmed, F., Zawad, S., Koch, F., Saad, W., and Boche, H. SafeCOMM: A Study on Safety Degradation in Fine-Tuned Telecom Large Language Models. *arXiv preprint arXiv:2506.00062*, 2025c.

Djuhera, A., Gacanin, H., and Boche, H. MambaCSP: Hybrid-Attention State Space Models for Hardware-Efficient Channel State Prediction, 2026a. URL https://arxiv.org/abs/2604.21957.

Djuhera, A., Koch, F., and Binotto, A. Joint Partitioning and Placement of Foundation Models for Real-Time Edge AI. In *2026 International Conference on Computing, Networking and Communications (ICNC)*, pp. 459–464. IEEE, 2026b.

GSMA. GSMA Open-Telco LLM Benchmarks 2.0. https://www.gsma.com/newsroom/article/gsma-open-telco-llm-benchmarks-ranks-frontier-ai-models-for-telco-ai-reveali

ng-critical-gap-in-network-automation-readiness/, 2025. Accessed: 2026-05-02.

He, Z., Zhang, X., Wang, Y., Lin, Y., Gui, G., and Gacanin, H. A Robust CSI-Based Wi-Fi Passive Sensing Method using Attention Mechanism Deep Learning. *IEEE Internet of Things Journal*, 10(19):17490–17499, 2023.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Hoydis, J., Cammerer, S., Ait Aoudia, F., Vem, A., Binder, N., Marcus, G., and Keller, A. Sionna: An Open-Source Library for Next-Generation Physical Layer Research. *arXiv preprint arXiv:2203.11854*, 2022.

Jiang, F., Pan, C., Dong, L., Wang, K., Debbah, M., Niyato, D., and Han, Z. A Comprehensive Survey of Large AI Models for Future Communications: Foundations, Applications and Challenges. *arXiv preprint arXiv:2505.03556*, 2025.

Jin, Y., Zhang, J., Jin, S., and Ai, B. Channel Estimation for Cell-Free mmWave Massive MIMO Through Deep Learning. *IEEE Transactions on Vehicular Technology*, 68(10):10325–10329, 2019.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020.

Karunaratne, S. and Gacanin, H. An Overview of Machine Learning Approaches in Wireless Mesh Networks. *IEEE Communications Magazine*, 57(4):102–108, 2019.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Li, Q., Sisk, P., Kannan, A., Yoo, T., Luo, T., Shah, G., Manjunath, B., Samarathungage, C., Boroujeni, M. T., Pezeshki, H., et al. Machine Learning Based Time Domain Millimeter-Wave Beam Prediction for 5G-Advanced And Beyond: Design, Analysis, and Over-the-Air Experiments. *IEEE Journal on Selected Areas in Communications*, 41(6):1787–1809, 2023.

Lin, X., Li, J., Baldemair, R., Cheng, T., Parkvall, S., Larsson, D. C., Koorapaty, H., Frenne, M., Falahati, S., Grovlen, A., and Werner, K. An Overview of the 3GPP Study on Artificial Intelligence for 5G New Radio. *arXiv preprint arXiv:2308.05315*, 2023.

Liu, B., Liu, X., Gao, S., Cheng, X., and Yang, L. LLM4CP: Adapting Large Language Models for Channel Prediction, 2024. URL https://arxiv.org/abs/2406.14440.

Liu, B., Gao, S., Liu, X., Cheng, X., and Yang, L. WiFo: Wireless Foundation Model for Channel Prediction, 2025a. URL https://arxiv.org/abs/2412.08908.

Liu, B., Lu, Y., Zhao, J., Yang, Q., Wu, W., Chen, L., Chauhan, J., and Zhang, J. WiLLM: an Open Framework for LLM Services over Wireless Systems. *arXiv preprint arXiv:2506.19030*, 2025b.

Maatouk, A., Ampudia, K. C., Ying, R., and Tassiulas, L. Tele-LLMs: A Series of Specialized Large Language Models for Telecommunications, 2024. URL https://arxiv.org/abs/2409.05314.

Maatouk, A., Ayed, F., Piovesan, N., De Domenico, A., Debbah, M., and Luo, Z.-Q. TeleQnA: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge. *IEEE Network*, 2025.

Masterman, T., Besen, S., Sawtell, M., and Chao, A. The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey. *arXiv preprint arXiv:2404.11584*, 2024.

Nikbakht, R., Benzaghta, M., and Geraci, G. TSpec-LLM: An Open-source Dataset for LLM Understanding of 3GPP Specifications, 2024. URL https://arxiv.org/abs/2406.01768.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in neural information processing systems*, 32, 2019.

Polese, M., Bonati, L., D'Oro, S., Johari, P., Villa, D., Velumani, S., Gangula, R., Tsampazi, M., Robinson, C. P., Gemmi, G., Lacava, A., Maxenti, S., Cheng, H., and Melodia, T. Colosseum: The Open RAN Digital Twin, 2024. URL https://arxiv.org/abs/2404.17317.

Saad, W., Hashash, O., Thomas, C. K., Chaccour, C., Debbah, M., Mandayam, N., and Han, Z. Artificial General Intelligence (AGI)-Native Wireless Systems: A Journey Beyond 6G. *Proceedings of the IEEE*, 2025. doi:10.1109/JPROC.2025.3526887.

Sheng, Y., Wang, J., Zhou, X., Liang, L., Ye, H., Jin, S., and Li, G. Y. A Wireless Foundation Model for Multi-Task Prediction, 2025. URL https://arxiv.org/abs/2507.05938.

Wang, L., Shelim, R., Saad, W., and Ramakrishnan, N. DMWM: Dual-Mind World Model with Long-Term Imagination, 2025a. URL https://arxiv.org/abs/2502.07591.

Wang, L., Shelim, R., Saad, W., and Ramakrishnan, N. World Model-Based Learning for Long-Term Age of Information Minimization in Vehicular Networks, 2025b. URL https://arxiv.org/abs/2505.01712.

Wang, X., Zhang, Y., and Cheng, N. BeamAgent: LLM-Aided MIMO Beamforming with Decoupled Intent Parsing and Alternating Optimization for Joint Site Selection and Precoding. *arXiv preprint arXiv:2603.18855*, 2026.

Xu, S., Kurisummoottil Thomas, C., Hashash, O., Muralidhar, N., Saad, W., and Ramakrishnan, N. Large Multi-Modal Models (LMMs) as Universal Foundation Models for AI-Native Wireless Systems. *IEEE Network*, 38(5): 10–20, 2024. doi: 10.1109/MNET.2024.3427313.

Yang, Z., Chi, G., Wu, C., Liu, H., Gao, Y., Liu, Y., Xu, J., and Han, T. X. Generative AI Meets Wireless Sensing: Towards Wireless Foundation Model, 2025. URL https://arxiv.org/abs/2509.15258.

Zhang, H., Ren, Y., Yuan, H., Zhang, J., and Shen, Y. Wi-Chat: Large Language Model Powered Wi-Fi Sensing. *arXiv preprint arXiv:2502.12421*, 2025a.

Zhang, K., He, H., Song, S., Zhang, J., and Letaief, K. B. Communication-Efficient Distributed On-Device LLM Inference over Wireless Networks. *IEEE Journal of Selected Topics in Signal Processing*, 2025b.

Zheng, J., Niyato, D., Zhao, C., Kang, J., and Wang, J. From Digital Twins to World Models:Opportunities, Challenges, and Applications for Mobile Edge General Intelligence, 2026. URL https://arxiv.org/abs/2603.17420.

Zou, H., Tian, Y., Wang, B., Bariah, L., Lasaulce, S., Huang, C., and Debbah, M. RF-GPT: Teaching AI to See the Wireless World, 2026a. URL https://arxiv.org/abs/2602.14833.

Zou, H., Yang, Y., Bariah, L., Tian, Y., Lu, Y., Wang, B., Bara, A., Mefgouda, B., Liu, H., Tao, Y., et al. Telecom World Models: Unifying Digital Twins, Foundation Models, and Predictive Planning for 6G. *arXiv preprint arXiv:2604.06882*, 2026b.

A. Related Work

Recent progress in AI has accelerated research across the wireless protocol stack. This includes both *AI for wireless*, where learning methods improve communication, sensing, and network-control tasks, and *wireless for AI*, where communication systems support distributed training, inference, and edge deployment of AI models. We briefly summarize the most relevant directions, with emphasis on foundation models, LLMs, agents, and world models.

A.1. AI for Wireless and Wireless for AI

Early AI-for-wireless research focused on well-scoped optimization and inference problems. Neural and reinforcement-learning methods have been applied to channel estimation (Jin et al., 2019), beam prediction (Li et al., 2023), resource allocation (Djigal et al., 2022), mesh networks (Karunaratne & Gacanin, 2019), trajectory optimization (Cheng et al., 2021), and sensing (He et al., 2023). These works demonstrated early that learning can be valuable when the task interface, input representation, and evaluation metric are clearly specified.

More recently, the rise of LLMs and attention-based architectures has motivated their adaptation to wireless sequence modeling. For example, Liu et al. (2024) and Djuhera et al. (2026a) adapt sequence models to channel-state prediction by designing CSI tokenization pipelines over time, frequency, and delay-domain representations. Beyond channel prediction, Zhang et al. (2025a) propose an LLM-powered Wi-Fi sensing system that maps raw CSI into textual or visual representations. Similarly, Wang et al. (2026) propose an LLM-aided MIMO beamforming framework in which the LLM parses natural-language intent into structured spatial constraints, while a dedicated numerical optimizer performs joint site selection and precoding. Furthermore, Zou et al. (2026a) introduce RF-GPT, an RF perception interface which converts IQ signals into spectrograms and grounds them through synthetic metadata and instruction tuning. These works show the potential of LLM-style interfaces, but they also reinforce a recurring pattern: successful systems either require task-specific tokenization or decouple the LLM from the underlying numerical wireless computation.

A complementary line of work studies *wireless for AI*, where communication systems are designed to support distributed learning and inference. This includes resilient federated and split-learning mechanisms under jamming, noise, or unreliable wireless links (Djuhera et al., 2025a;b), as well as joint model partitioning, scheduling, and inference placement across edge servers, MEC nodes, and on-device LLM services (Djuhera et al., 2026b; Zhang et al., 2025b; Liu et al., 2025b). These efforts highlight that AI-native wireless is not only about applying AI to the network, but also about adapting the network to the computational and communication

demands of modern AI workloads.

More recently, related trends seek to reproduce the cross-task capabilities of LLMs through telecom-specific post-training. For example, Nikbakht et al. (2024); Maatouk et al. (2024; 2025) develop telecom-oriented datasets and models for standards understanding, Q&A, retrieval, and domain-specific reasoning. These efforts are important first steps, but they mostly train on textual artifacts of telecom knowledge rather than on the operational substrate of wireless systems. They therefore improve language-level expertise, but do not by themselves ground models in RF measurements, network dynamics, control actions, or deployment feedback. To bridge this gap, recent work has proposed large wireless foundation models, multimodal wireless models, and wireless world models (Alikhani et al., 2024; Xu et al., 2024; Zou et al., 2026b; Saad et al., 2025). However, many of these proposals remain at the level of architectural vision, with limited discussion of the data interfaces, provenance, operational feedback, and evaluation protocols required for deployment. This raises the question of whether current foundation and world model visions adequately address the structural data bottlenecks discussed in Sec. 3.

Overall, existing work validates the usefulness of both classical neural methods and attention-based models for wireless applications. However, most successes remain tied to scoped tasks, carefully designed representations, and explicit optimization or system interfaces. This motivates our central position: deployable AI-native wireless should compose specialized models, solvers, simulators, and reasoning agents, rather than assuming that a single monolithic model can absorb the entire protocol stack.

A.2. Simulators, Datasets, and Benchmarks

As AI-based methods become increasingly common in wireless research, high-quality simulators and experimental platforms are essential for data generation, training, testing, and reproducible evaluation. Platforms such as DeepMIMO (Alkhateeb, 2019), Sionna (Hoydis et al., 2022), OpenRAN Gym (Bonati et al., 2023), and Colosseum (Polese et al., 2024) provide important infrastructure for simulating wireless environments, signal propagation, protocol behavior, and network-control loops, often with interfaces to modern AI frameworks such as PyTorch (Paszke et al., 2019). These platforms have been instrumental for developing and validating AI-based wireless models. At the same time, they also encode assumptions about channel models, propagation, traffic, mobility, hardware impairments, and protocol abstractions. As a result, models trained in one simulation environment may learn simulator-specific artifacts rather than transferable wireless structure. This motivates systematic cross-simulator and sim-to-real studies, as well as stronger documentation of simulator provenance, calibration, and standards compliance.

Beyond simulators, recent works have introduced telecom-oriented datasets for LLM post-training and evaluation (Nikbakht et al., 2024; Maatouk et al., 2024; 2025). As discussed above, these datasets are useful for standards understanding, Q&A, and retrieval, but remain predominantly text-based and therefore provide limited grounding for wireless use cases. They also inherit common risks of post-training, including safety degradation and domain overfitting. For instance, Djuhera et al. (2025c) show that telecom post-training can degrade safety behavior and introduce TeleHarm as a telecom-specific safety realignment dataset. These findings suggest that telecom datasets must move beyond textual knowledge artifacts toward multi-modal, configuration-aware, and operationally grounded data.

Benchmarking remains similarly underdeveloped. Current telecom LLM benchmarks, including the benchmarks proposed by GSMA (2025), mostly measure standards knowledge, Q&A, configuration generation, or troubleshooting, but do not yet capture closed-loop control quality, causal intervention accuracy, robustness under distribution shift, or deployment safety. This raises a central question: *How should actionable wireless intelligence be measured?* A credible benchmark suite would need to evaluate not only language-level correctness, but also whether a system can retrieve the right state, choose the right tool, respect protocol and safety constraints, and improve network KPIs under realistic operational conditions. Without such benchmarks, wireless world models may remain compelling architectural visions, but their practical value for deployable network intelligence cannot be meaningfully assessed.

While a complete survey is beyond the scope of this paper, existing trends show that both classical neural methods and attention-based models are increasingly used to improve wireless applications. However, as argued in the main body, most works remain tied to scoped tasks, custom representations, and isolated benchmarks, leaving open the question of how these components should be composed into deployable, auditable, and configuration-aware AI-native wireless systems that are compatible with existing telecom infrastructure, protocol stacks, and standards.

B. The Wireless Data Problem and Data Cards

Wireless systems generate enormous amounts of data, but much of it is difficult to reuse for general-purpose AI. Compared with text, code, or images, an IQ stream, CSI tensor, KPI trace, or scheduler log is meaningful only relative to its configuration, protocol state, measurement setup, simulator assumptions, and deployment context. In Sec. 3, we identified four structural bottlenecks that make monolithic wireless models, including wireless world models, difficult

to develop and validate. In particular, the lack of a universal documentation scheme prevents wireless data from becoming a reusable foundation-model substrate.

This motivates our composable and agentic view of AI-native NextG: rather than forcing all wireless data into a single model, specialized components should consume well-documented data through explicit interfaces compatible with existing protocol and standardization boundaries. However, agentic wireless AI still requires a substantial redesign of how wireless datasets are categorized, documented, and shared. We therefore propose *wireless data cards*: structured documentation artifacts that record the configuration, provenance, task, and deployment assumptions behind each dataset. This is inspired by dataset and model documentation practices in responsible ML, but adapted to the particular needs of wireless systems.

A minimal wireless data card should record:

- **Task:** prediction target, action space, loss function, latency budget, and safety constraints.
- **Configuration:** carrier frequency, bandwidth, numerology, antenna geometry, beam codebook, pilot design, and scheduler settings.
- **Provenance:** measurement source or simulator, simulator version, channel model, geometry, mobility and traffic models, impairment assumptions, and random seeds.
- **Environment:** indoor/outdoor setting, topology, blockage, deployment density, UE distribution, and mobility.
- **Protocol Layer and Timescale:** PHY, MAC, RAN, core, edge, application layer and interaction frequency.
- **Policy:** operator objectives such as fairness, energy, QoS/QoE, as well as regulatory and safety constraints.
- **Evaluation:** baselines, distribution shifts, failure criteria, uncertainty calibration, and robustness tests.

Without such documentation, wireless datasets remain configuration-opaque and difficult to compose across tasks, simulators, and deployments. For agentic systems, data cards are even more important: they allow an agent to decide which dataset, simulator, model, or tool is valid for a given configuration and control objective. Future extensions may also include reasoning traces, tool-call histories, simulator rollouts, and operator feedback, enabling agentic workflows to be audited and improved over time. The call for wireless data cards is therefore a call for wireless AI to adopt the same rigor in data documentation that modern AI already demands, while accounting for the configuration- and deployment-dependence unique to wireless systems.

C. Wireless World Models

In this section, we formalize the notion of a wireless world model and review related work.

C.1. Definition and Taxonomy

World models are appealing because they promise prediction before action. In their classical form, a world model learns an action-conditioned transition model,

$$p_\theta(s_{t+1:t+H} \mid s_t, a_{t:t+H}), \tag{1}$$

where s_t denotes the current latent or observed system state, $a_{t:t+H}$ denotes a candidate action sequence over a planning horizon H, and $s_{t+1:t+H}$ denotes the predicted future trajectory under that action sequence. Such a model allows an agent to imagine possible futures, evaluate candidate actions, and plan under uncertainty. However, in the LeCun sense (Assran et al., 2023), a world model is useful only when the "world", the state representation, the action space, and the prediction targets are well defined. For wireless systems, this immediately raises a specification problem. Is the state an electromagnetic field, a CSI tensor, a queue vector, a traffic map, a slice state, a protocol trace, an operator policy, or all of these jointly? Are the actions beam updates, scheduling decisions, handover parameters, routing rules, slice budgets, energy-saving policies, or application-level intents? What is the prediction horizon: microseconds, slots, seconds, minutes, or hours? What objective should be optimized: throughput, latency, reliability, energy, fairness, SLA compliance, or some operator-specific trade-off? Without specifying these elements, a wireless world model is not yet a model, but only a metaphor.

C.2. Recent Work on Wireless World Models

Recent wireless and telecom world model proposals are most convincing precisely when they avoid ambiguity.

Initial work on world model-based learning for long-term age-of-information (AoI) minimization defines a concrete mmWave V2X link-scheduling problem, a packet-completeness-aware AoI objective, a recurrent state-space model (RSSM) for latent dynamics prediction, and a Sionna-based simulator with ray tracing and vehicle mobility (Wang et al., 2025b). The follow-up dual-mind wireless world model (Wang et al., 2025a) further specializes this setting by adding a pattern-driven System 1 component and a logic-driven System 2 component, aiming to improve long-horizon imagination and scheduling consistency under mobility, blockage, and link-availability constraints. This line of work is valuable because the "world" is carefully scoped: the state contains packet-completeness-aware AoI (CAoI), vehicle locations, and channel/ray-tracing features, the action is link scheduling, and the objective is long-horizon

CAoI minimization. It is therefore a useful local world model, but not evidence that a single model can absorb the entire wireless stack.

At the network level, Zou et al. (2026b) provide one of the most explicit attempts to define a telecom world model (TWM). Rather than treating the world as an undifferentiated wireless latent space, TWM decomposes the telecom system into a controllable system world and an exogenous world. It then introduces three layers: a *Field World Model* for spatial and electromagnetic prediction, a *Control/Dynamics World Model* for action-conditioned KPI trajectories, and a *TelecomGPT layer* for intent translation, tool orchestration, and guardrail enforcement. This formulation explicitly clarifies what a telecom world model would need to provide: state grounding, action-conditioned rollouts, uncertainty estimates, multi-timescale dynamics, model-based planning, and safety-aware interaction with operators. However, it also reinforces our main argument. The strongest TWM formulation is not monolithic, but decomposed, tool-integrated, simulator-dependent, and explicitly agentic at the foundation model layer.

In addition, recent work connecting digital twins and world models (Zheng et al., 2026) argues that learned predictive models can accelerate simulation, support counterfactual planning, and enable mobile edge or network intelligence beyond reactive control. We agree with this direction when world models are treated as scoped predictive components. Digital twins provide high-fidelity mirroring, validation, and data generation, while world models may provide faster learned rollouts for specific control loops. The problem arises when this complementarity is replaced by the stronger claim that a single wireless world model should become the central substrate for AI-native networking.

C.3. Structural Data Bottlenecks Remain

As argued in the main body, wireless world models inherit the same data bottlenecks as wireless LLMs, but in a stricter form. A world model must predict transitions under actions. Missing context therefore does not merely reduce accuracy, but invalidates counterfactual rollouts. Configuration dependence becomes a transition-model problem: the same CSI tensor, beam update, or scheduler action can imply different future dynamics under different numerologies, antenna layouts, traffic models, or beam codebooks. Simulation dependence becomes a calibration problem: if rollouts are generated by a simulator with biased propagation, mobility, or protocol assumptions, the world model may learn simulator artifacts as if they were wireless physics. The absence of a universal wireless token becomes an interface problem: IQ samples, CSI tensors, topology graphs, KPI logs, and operator intents cannot be fused by simply scaling a single sequence model without destroying useful abstraction boundaries. Finally, missing operational feedback becomes a causal-identifiability problem: historical telemetry reflects actions chosen by existing operator policies, not the counterfactual outcomes of unseen interventions. Thus, world models do not bypass the wireless data problem and make configuration metadata, simulator provenance, cross-layer interfaces, and operational calibration even more load-bearing.

Our argument is therefore not that wireless world models are useless. Local world models can be useful for scoped prediction problems: radio-map forecasting, traffic evolution, slicing rollouts, failure propagation, link scheduling, or counterfactual policy evaluation. The problem is the stronger claim that monolithic or single wireless world models should become the primary substrate of AI-native networking. In the wireless domain, the relevant world is fragmented across PHY, MAC, RAN, transport, core, edge, and application layers, each with different observations, actions, timescales, constraints, and safety requirements. The more realistic path is to treat world models as specialized predictive tools inside a broader agentic architecture, alongside classical algorithms, RF perception models, digital twins, telemetry databases, standards-aware retrieval, and safety monitors. World models are thus useful only when the world is sufficiently scoped, especially in wireless which is a hierarchy of coupled worlds operating at incompatible timescales.

D. Agentic Wireless Network Architectures

In this section, we expand on the agentic wireless network architecture proposed in Sec. 5 and summarize why the limitations of current wireless LLMs and world models point toward such composable architectures. Furthermore, we outline how agentic reasoning can be integrated with existing telecom infrastructure, protocol stacks, and workflows.

D.1. Why the Evidence Points to Agentic Architectures

The case for agentic wireless AI follows directly from the structural bottlenecks discussed in Sec. 3. Most importantly, the lack of a universal wireless token actively prohibits training truly multimodal wireless foundation models. Accordingly, monolithic models cannot resolve these bottlenecks, and we argue that only agentic systems can overcome them.

For example, a reasoning agent can query configuration databases, retrieve standards constraints, call specialized PHY/MAC/RAN models, invoke classical solvers, run digital twin validations, and check safety policies before recommending an action. However, this does not imply that AI-based methods should replace existing wireless algorithms. Many signal processing, estimation, optimization, and scheduling methods remain preferable in regimes where they are reliable, interpretable, and computationally effi-

cient. The role of the agent is therefore not to replace the wireless stack, but to select, compose, and validate the right component for the current task, configuration, and risk level.

This view is also consistent with how telecom systems are standardized and deployed. 3GPP and O-RAN do not expose the network as one end-to-end differentiable system, but instead define functions, interfaces, measurements, control loops, and constraints. Agentic wireless AI treats this modularity as a design advantage, whereas monolithic models tend to obscure it. Rather than asking one model to internalize the entire protocol stack, agentic introduces intelligence at the interface level, where reasoning, retrieval, tool use, validation, and human oversight can be made explicit.

D.2. Architecture: Agents Around the Protocol Stack

An agentic wireless network architecture consists of three layers connected by an *agent harness* (see Fig. 2). The first layer is the *protocol and infrastructure layer*, including PHY/MAC procedures, RAN control, transport/core functions, telemetry pipelines, configuration databases, and existing management systems. The second layer is the *tool layer*, exposing specialized components through callable interfaces: channel estimators, beam predictors, interference classifiers, schedulers, optimizers, simulators, digital twins, standards retrieval, and safety monitors. The third layer is the *reasoning layer*, where one or more agents plan workflows, select tools, compare outputs, validate constraints, and communicate with operators. The harness connects these layers by managing context, permissions, tool schemas, logging, guardrails, rollback, and human approval.

In a practical deployment, multiple agents can operate at different scopes and timescales. A PHY-facing agent may assist with model selection, channel estimator validation, or beam management diagnosis. A RAN-control agent may coordinate xApps/rApps for load balancing, energy saving, slicing, or mobility optimization. A safety agent can check proposed actions against SLA, regulatory, vendor, and rollback constraints. The agents coordinate through structured tool calls and shared state summaries, while latency-critical decisions remain inside specialized control loops.

This architecture is compatible with current infrastructure. O-RAN already separates control across near-real-time and non-real-time timescales and exposes modular xApps/rApps, telemetry streams, and policy interfaces. Containerized deployment further allows operators to add, update, or roll back individual models and tools without retraining a full-stack model. Agent skills can be defined as bounded capabilities, such as `query_kpi`, `retrieve_3gpp_constraint`, or `compare_estimators`. This makes the system highly maintainable, as operators can audit which tools were called, which assumptions were used, which constraints were checked, and why a recommendation was made.

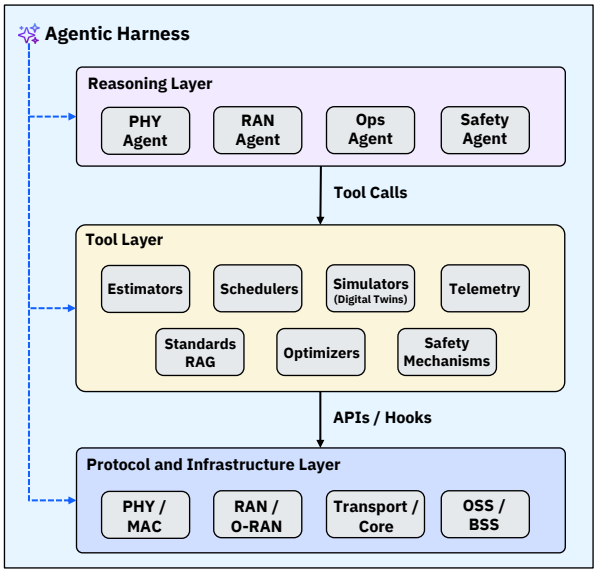

Figure 2. Three-layer agent harness for wireless networks.

D.3. Example Workflows

We sketch three representative workflows that illustrate how agentic wireless AI composes existing components.

Model Selection for Channel Estimation. A gNB observes degraded CSI quality for a subset of UEs under changing mobility and SNR conditions. The agent retrieves the carrier configuration, pilot pattern, antenna geometry, UE mobility profile, recent CSI error statistics, and historical channel estimation performance. It then compares a classical LS/MMSE estimator, a lightweight neural estimator, and a larger learned CSI predictor under the current configuration. The candidates are evaluated on recent telemetry and, when needed, on simulated channel realizations with matched numerology and mobility. If the neural estimator improves NMSE while satisfying latency and uncertainty constraints, the agent recommends enabling it. If the scenario is outside the learned model's documented validity range, the agent falls back to the classical estimator and flags the need for additional data. Thus, the estimator is selected per configuration and risk level, not replaced globally.

Scheduler Adaptation Under Slice Pressure. A cell experiences rising URLLC delay while eMBB throughput remains high. The agent retrieves per-slice queues, PRB utilization, HARQ statistics, traffic forecasts, current scheduler weights, and SLA constraints. It then calls a queueing-based baseline, a learned scheduler surrogate, and a digital twin rollout to compare candidate scheduling policies. Unsafe policies are rejected if they violate URLLC latency, starve eMBB traffic, or exceed predefined fairness limits.

The agent may recommend a bounded update to scheduler weights, a temporary slice-priority adjustment, or no change if the evidence indicates transient congestion. The key point is that the agent does not invent a scheduler from scratch. It selects and validates among existing scheduling mechanisms using telemetry, prediction, and policy constraints.

Intent-to-Configuration for Network Operations. An operator issues a high-level intent such as *"Prepare this region for a stadium event while preserving URLLC reliability and limiting energy cost."* The agent retrieves the relevant cells, neighbor relations, traffic history, mobility forecasts, slice policies, and applicable 3GPP/O-RAN constraints. It decomposes the intent into candidate actions, such as temporary slice budget changes, mobility parameter updates, admission control thresholds, or activation of additional carriers. Each candidate is checked against standards constraints, evaluated through a digital twin or learned surrogate, and passed through a safety monitor before being surfaced to the operator. The final output is not a raw configuration dump, but a ranked plan with assumptions, expected KPI impact, uncertainty, and rollback conditions. This illustrates the appropriate role of foundation models for translating intent and orchestrating tools, while validated components perform the wireless computation.

These examples illustrate the key architectural principle: intelligence arises from composition. The agent does not hallucinate network states when telemetry can be queried, does not reinvent protocol behavior when standards can be retrieved, and does not replace a validated solver when a solver is the right tool. Such workflows can be learned from wireless reasoning traces, which are already implicit in current network operations through telemetry queries, configuration changes, troubleshooting tickets, simulator runs, and operator decisions. Agentic wireless AI is therefore not a rejection of foundation models, world models, or classical algorithms. It is a deployment-oriented framework for deciding when each of them should be used.