5G/6G Academy
TELECOM CERTIFICATIONS · SINCE 2009

Technical Whitepaper
N° 04 · June 2026

TECHNICAL WHITEPAPER

The AI-Native RAN: What's Standardized, What's Deployed

AI entered the radio network three times — through the core, the management plane, and finally the air interface itself. This paper separates what 3GPP has actually specified from what vendors demonstrate, what operators run today, and what "AI-native 6G" can credibly mean.

AUTHOR
5G/6G Academy Research

VERSION
1.0

DATE
June 2026

TR 38.843 | REL-19 · REL-20 | RIC & NWDAF | AI-RAN ALLIANCE | 6G

5g6gacademy.com/whitepapers

5G/6G ACADEMY · TECHNICAL WHITEPAPER · N° 04 June 2026 · 5g6gacademy.com

The AI-Native RAN: What's Standardized, What's Deployed

5G/6G Academy Research

a TELCOMA Global publication · Telecom Certifications · Since 2009

***Abstract*—"AI-native" is 2026's most-used and least-defined RAN phrase. The standardized reality is narrower and more interesting: 3GPP's Rel-18 study (TR 38.843) mapped three air-interface use cases; Rel-19 made the one-sided ones — beam prediction and positioning — normative, together with a lifecycle-management framework; and in June 2025 Rel-20 took the hard one, two-sided CSI compression, into normative work that includes a standardized reference encoder and a RAN4 test decoder — quietly bending 3GPP's own rule that models stay out of specifications. Meanwhile the only operator-named deployment numbers are about electricity: an Ericsson-Vodafone London trial reported up to 33% daily radio power savings, self-measured. We map the three entry points of AI into the network, separate 3GPP's AI-for-RAN from the AI-RAN Alliance's GPU-convergence agenda, and close with a sober reading of the 6G "AI-native air interface" — where the rate limiter may be type approval, not neural networks.**

1. Introduction: Three Doors Into the Network

Artificial intelligence did not enter the mobile network through the radio. It entered through the billing-adjacent, safety-distant layers first — the core's analytics function in Rel-15, the management plane's optimization loops — and only reached the air interface, the most standardized and most real-time part of the system, in the last two releases. That ordering was not an accident. A mispredicted analytics report wastes a dashboard; a mispredicted beam drops your call. The closer AI gets to the radio frame, the higher the cost of being wrong, and the more machinery you need for monitoring, fallback and interoperability.

This paper follows AI through its three doors. The core door: NWDAF, the network data analytics function, now a decade old in spec terms. The management door: RAN3's data-collection and optimization work, where energy saving quietly became AI's first production success. And the air-interface door: the Rel-18 study, the Rel-19 normative step, and the Rel-20 work item that takes on the hardest problem in the set. Along the way we keep two things rigorously separate that marketing tends to blur: 3GPP's **AI for the RAN** — models improving radio functions — and the AI-RAN Alliance's **AI and the RAN** — radio and AI workloads sharing the same GPU infrastructure. They are different stories with different economics, and conflating them is the most common error in current coverage.

2. The Map Rel-18 Drew

The first 3GPP study of AI/ML for the NR air interface — FS_NR_AIML_air, launched from study item RP-221348 and reported in TR 38.843 — deliberately picked three pilot use cases rather than boiling the ocean: CSI feedback, beam management, and positioning [1], [3]. Table 1 summarizes what each model consumes, what it produces, and the property that turned out to decide everything: how many *sides* of the radio link the model touches.

TABLE 1 · The three air-interface use cases — and where their models live

Use case	The model sees…	…and produces	Sides	Status
Beam management	measurements of a small subset of beams (spatial) or of past beams (temporal)	a prediction of the best beam — without sweeping them all	one-sided (UE- or network-side)	normative in Rel-19 [4]
Positioning	channel measurements / RF fingerprints	a location directly, or an improved measurement for the classic chain	one-sided (UE, gNB or LMF; Cases 1, 3a, 3b prioritized)	normative in Rel-19 [4]
CSI feedback	the downlink channel state	a compressed report the gNB reconstructs	two-sided — UE encoder, gNB decoder	study through Rel-19 → normative in Rel-20 (Case 0) [5]

Beam prediction is the intuitive winner: instead of sweeping every candidate beam and reporting them all, a model watches a subset — or a history — and predicts where the energy should go. Positioning came in two flavors: *direct* models that output a location (fingerprinting-style) and *assisted* models that output a better measurement for the existing positioning chain [1]. Both are **one-sided**: the model lives entirely at one end, so nobody else's equipment needs to understand its internals.

CSI compression is the odd one out — an autoencoder split across the air gap, encoder in the phone, decoder in the base station, "with a tight interplay between the UE and gNB" in 3GPP's own words [1]. That single architectural fact delayed it two releases, as Section 4 explains.

The study also set the scoping rule that has governed everything since: the models themselves are, in 3GPP's careful phrasing, "not expected to be specified" — specifications cover the plumbing instead: data collection, signaling, capability indication, and lifecycle management (LCM) with monitoring, activation, deactivation and fallback [1]. Standardize the hooks, compete on the models.

3. Rel-19: AI Crosses Into the Air Interface

At RAN#102 in Edinburgh (December 2023), 3GPP approved the Rel-19 work item — WID RP-234039 — that made the one-sided use cases normative: beam prediction in both its spatial and temporal variants (BM-Case1 and BM-Case2, supporting both UE-side and network-side models) and AI/ML positioning, with first priority on the UE-based, gNB-assisted and LMF-side configurations (Cases 1, 3a and 3b), all wrapped in a general LCM framework for one-sided models [4], [6]. From Rel-19 onward, "AI in the air interface" is no longer a research statement; it is signaling a phone can declare support for.

The network side moved in parallel. Building on the Rel-17 framework study (TR 37.817) and the Rel-18 work item that covered network energy saving, load balancing and mobility optimization — RAN3 closed its Rel-18 normative work in the fourth quarter of 2023 — the Rel-19 work item specified AI/ML-assisted **network slicing** (predicted slice-level information in data collection, per-slice UE performance feedback) and **coverage-and-capacity optimization**, adding future CCO states and predicted CCO issues to the Xn and F1 interfaces [2]. For Rel-20, RAN endorsed at RAN#106 (Madrid, December 2024) that NG-RAN AI/ML work stays on the current 5G architecture and interfaces — evolution, not reinvention [2].

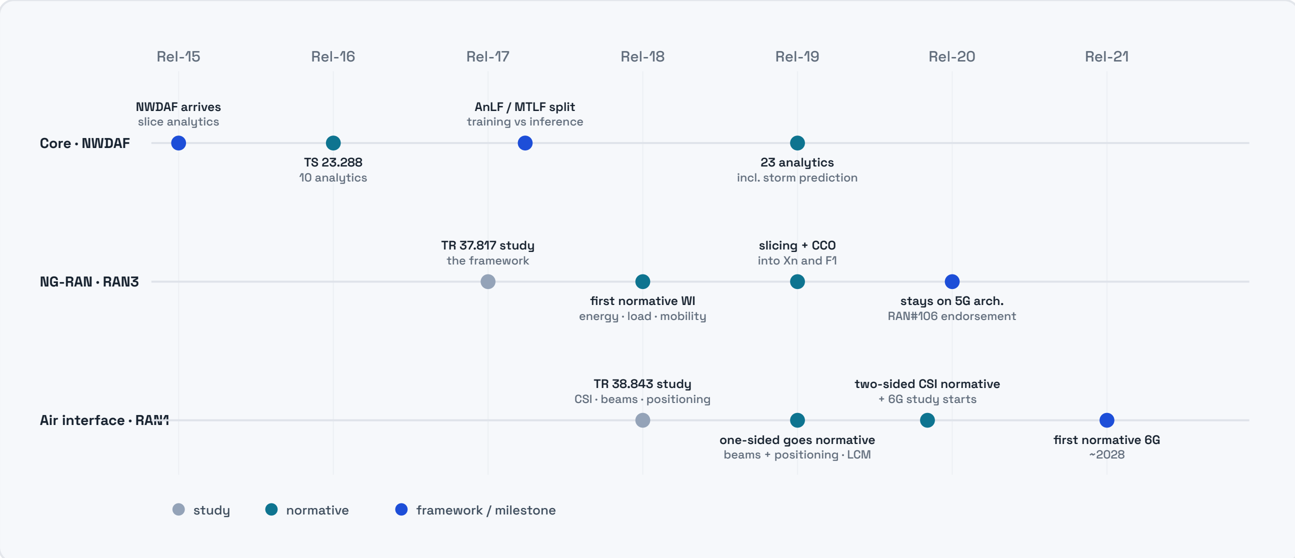

Fig. 1. Three doors, one direction: AI/ML standardization across the core (NWDAF), the NG-RAN (RAN3) and the air interface (RAN1), Rel-15 to Rel-21. Gray = study; teal = normative; blue = framework milestones. The air interface was last through the door — and is moving fastest now. Sources: [1]–[7], [10].

For completeness, the oldest door: **NWDAF** arrived in Rel-15 for slice analytics, got its own specification (TS 23.288) in Rel-16, and split in Rel-17 into an inference function (AnLF) and a model-training function (MTLF). Its standardized analytics grew from 10 in Rel-16 to 23 by Rel-19 — counting by analytics identifiers in TS 23.288, a tally that includes signalling-storm prediction and QoS/policy assistance in the latest release [9], [10]. The core has been quietly accumulating AI surface for seven years; the air interface is catching up in two.

4. The Two-Sided Frontier

Why did CSI compression — the use case with the largest theoretical payoff — take until Rel-20? Because a two-sided model is an interoperability problem disguised as a machine-learning problem. The encoder is trained by whoever builds the phone's modem; the decoder by whoever builds the base station. For the pair to work, they must be trained *together*, or against something common — and "inter-vendor training collaboration," in the Rel-19 WID's own words, was exactly the unresolved issue that kept CSI compression in study while its one-sided siblings went normative [4].

At RAN#108 (Prague, June 2025) 3GPP approved the Rel-20 work item — acronym NR_AIML_air_Ph2 — that takes two-sided CSI compression normative, starting with the baseline "Case 0": spatial-frequency compression of current-slot CSI, UE encoder, gNB decoder [5], [7]. And here is the quiet doctrinal shift: to make vendor A's encoder meet vendor B's decoder, the work item brings a **standardized reference model structure for the encoder and a RAN4 test decoder** into the specification orbit [5], [8]. The "models are never standardized" rule, stated almost as scripture since

2022, now has its first carefully-fenced exception. How far that fence moves is, in our view, the single most consequential open question in RAN standardization: it decides whether AI-native interfaces stay multi-vendor or quietly become a lock-in mechanism.

5. Meanwhile, In Production: AI Came Through the Electricity Bill

Strip away the demos and ask what AI is verifiably doing in commercial networks today, with an operator's name attached, and the answer is: saving power. In March 2025, Vodafone UK and Ericsson reported a London trial of an AI app suite — predictive "deep sleep" hibernation, behaviorally-tuned 4G cell sleep, and an ML heatmap ranking underperforming sites — that cut daily 5G radio-unit power consumption by up to 33% at selected sites, with deep sleep reaching up to 70% savings in low-traffic hours [11]. The honest labels matter: it was a trial, at selected sites, self-measured by the vendor and operator, and "up to" is doing real work in those sentences. It is still the most concrete operator-named AI-in-the-RAN result on the public record.

Energy is the perfect Trojan horse for RAN AI, and it is worth being precise about why: the downside is bounded (worst case, a cell sleeps too eagerly and wakes on demand), the payoff lands directly on the operating bill, and no handset needs to know anything happened. Contrast that with the air-interface features of Section 3: as of this writing — mid-2026 — we could not verify **any** operator running standardized Rel-19 air-interface AI features in commercial service. The features are simply too new; silicon and network software carrying them are still working through the pipeline. The deployed reality of "AI in the RAN" in 2026 is network-side optimization and automation apps — the rApp story our Open RAN paper documents — not neural networks in the radio frame [12].

6. The Other "AI-RAN": Sharing the Computer

A second, entirely different agenda trades under a confusingly similar name. The **AI-RAN Alliance** — founded at MWC Barcelona on February 26, 2024 by eleven companies including NVIDIA, SoftBank, Ericsson, Nokia, Samsung and T-Mobile — organizes its work into three pillars: *AI-for-RAN* (models improving the radio, the 3GPP-adjacent part), *AI-and-RAN* (radio and AI workloads sharing the same accelerated infrastructure), and *AI-on-RAN* (AI services delivered over the network) [13]. By February 2026 membership had reached 132 [14]. The center of gravity is the middle pillar — the proposition that the same GPU estate can run your DU today and rent out inference capacity tonight.

The flagship is SoftBank's AITRAS: a converged stack running carrier-grade vRAN and AI workloads on NVIDIA Grace/GH200 infrastructure, with SoftBank claiming its Grace-based CU doubles the number of radio units one CU server can host — a vendor self-claim, unbenchmarked publicly [15]. The program's trajectory has firmed up: field deployment from FY2025, an orchestrator open-sourced in February 2026, and commercial availability targeted for October 2026, starting inside the SoftBank group, with expansion to global operators "from 2026" onward [15], [16]. Whether the economics of GPU-RAN convergence survive contact with depreciation schedules is a question for accountants as much as engineers — but it is a *different question* from whether AI improves the radio, and readers should refuse any slide that mixes the two.

7. "AI-Native 6G": Claim Versus Evidence

The 6G framework makes AI official: ITU-R M.2160 lists "Artificial Intelligence and Communication" among the six IMT-2030 usage scenarios, with ubiquitous intelligence named among the framework's overarching design aspects [17]. On the research side, the ambition is concrete: replace designed signal-processing blocks with learned ones. Nokia Bell Labs' DeepRx demonstrated a full neural receiver — frequency-domain samples in, bits out — performing at or beyond conventional receivers on 3GPP channel models [18], and the broader "AI-native air interface" program imagines learned waveforms and transceiver pairs co-designed end to end [19].

The sober counterweight comes from inside the same industry. A 2024 IEEE ComSoc piece — authored from Nokia and York University — lists the unsolved engineering: models that generalize across radio conditions instead of overfitting the training campaign; lifecycle management with real monitoring and graceful fallback to non-AI operation; the impossibility of guaranteeing hard constraints in a neural network; black-box opacity; and the compute and energy cost of running large models in places engineered for watts [20]. To that list we add one of our own: **type approval**. Every radio feature today is conformance-tested against deterministic expected behavior; nobody has yet demonstrated how to certify a learned receiver whose behavior is a function of its training data. It is telling that the first thing Rel-20's two-sided work standardized was a *test decoder* [5] — the certification machinery is being invented alongside the feature. Our reading: the rate limiter for AI-native 6G is less likely to be neural-network capability than the institutional machinery of proving, to a regulator and to a roaming partner, that the learned thing behaves.

8. What To Do Now

For engineers and planners, the 2026 state of play converts into four concrete moves. **Build the data pipeline before the models**: every standardized AI feature, from NWDAF analytics to Rel-19 beam prediction, is fed by the data-collection machinery — that plumbing is vendor-neutral, useful immediately, and required later. **Pilot where the downside is bounded**: energy-saving apps and rApp-class automation are production-proven territory with measurable OPEX returns; treat air-interface AI as a 2027 device-cycle conversation. **Ask vendors the two-sided question**: whether their CSI-compression roadmap targets the standardized reference model path or proprietary pairing — the answer reveals their lock-in strategy more honestly than any slide. And **watch four markers**: the first commercial activation of a Rel-19 air-interface feature; the progress of NR_AIML_air_Ph2 toward its reference model; whether the AI-RAN Alliance's convergence economics produce an audited deployment; and the first RAN4 conformance regime for a learned feature — the moment "AI-native" stops being a slogan and becomes a certified property.

AI is in the RAN to stay. But in 2026 the truthful sentence is narrower than the billboards: the hooks are standardized, the one-sided features are specified, the two-sided frontier is being negotiated, and the only audited wins are on the electricity bill. That is not a disappointment — it is exactly how the air interface, the most carefully engineered radio system humans operate, was always going to adopt machine learning: monitoring first, fallback always, models last.

References

[1] 3GPP, "AI/ML for NR Air Interface" (rapporteur article, J. Montojo), 3GPP Technology Highlights. `3gpp.org/technologies/ai-ml-nr`

[2] 3GPP, "AI/ML for NG-RAN & 5G-Advanced towards 6G" (RAN3 leadership article), Jun. 2025. 3gpp.org/technologies/ai-ml-ngran

[3] 3GPP, RP-221348, "Revised SID: Study on AI/ML for NR air interface," TSG RAN#96, Jun. 2022 (output: TR 38.843). 3gpp.org file server

[4] 3GPP, RP-234039, "New WID on AI/ML for NR air interface," TSG RAN#102, Edinburgh, Dec. 2023. 3gpp.org file server

[5] 3GPP, RP-251870, "New WID: AI/ML for NR air interface enhancements (NR_AIML_air_Ph2)," TSG RAN#108, Prague, Jun. 2025. 3gpp.org file server

[6] 5G-PPP / 6GStart, "Summary report from 3GPP plenaries #102 (Edinburgh, Dec. 2023)." standards-tracker.5g-ppp.eu

[7] Qualcomm, "3GPP Release 20: completing the 5G-Advanced evolution, preparing for global 6G standardization," OnQ blog, Jun. 2025. qualcomm.com/news/onq

[8] A. et al., "AI/ML life cycle management for interoperable AI-native RAN," arXiv:2507.18538, 2025. arxiv.org/abs/2507.18538

[9] 3GPP, TS 23.288, "Architecture enhancements for 5G System to support network data analytics services" (analytics IDs recounted: 10 in v16, 23 in v19). portal.3gpp.org

[10] Anon., "Towards network data analytics in 5G systems and beyond," arXiv:2506.04860, 2025. arxiv.org/abs/2506.04860

[11] Ericsson, "Vodafone UK and Ericsson trial AI solutions for improved 5G energy efficiency," press release, Mar. 11, 2025. ericsson.com/en/press-releases/3/2025

[12] 5G/6G Academy, "Open RAN in Practice: What the 2026 Deployments Actually Tell Us," Whitepaper Nº02, Jun. 2026. 5g6gacademy.com/whitepapers

[13] MWC Barcelona, "Industry leaders in AI and wireless form AI-RAN Alliance," Feb. 26, 2024. mwcbarcelona.com

[14] AI-RAN Alliance, "MWC 2026 momentum" (132 members), Feb. 2026. ai-ran.org/press-releases

[15] SoftBank Corp., "AITRAS implements centralized and distributed AI-RAN architectures using Arm-based NVIDIA platform," press release, Mar. 3, 2025. softbank.jp/en/corp/news/press

[16] SoftBank Corp. / Businesswire, "SoftBank Corp. announces development of AITRAS," Nov. 12, 2024. businesswire.com

[17] ITU-R, Recommendation M.2160-0, "Framework and overall objectives of the future development of IMT for 2030 and beyond," Nov. 2023. itu.int

[18] M. Honkala, D. Korpi, J. M. J. Huttunen, "DeepRx: Fully convolutional deep learning receiver," arXiv:2005.01494, 2020. arxiv.org/abs/2005.01494

[19] J. Hoydis, F. Ait Aoudia, A. Valcarce, H. Viswanathan, "Toward a 6G AI-native air interface," arXiv:2012.08285, Nokia Bell Labs, 2020. arxiv.org/abs/2012.08285

[20] IEEE ComSoc CTN, "AI native in wireless radio access networks: is there a pot at the end of this rainbow?," Oct. 2024. comsoc.org/publications/ctn