AI-NATIVE 6G AND AI-RAN

September 2026 evidence update

AI-Native 6G and AI-RAN: Standards Convergence, Verification and Diverging Compute Paths

A longitudinal reading from the May Apex Standards baseline through Singapore and Madrid: what is established, what is beginning to harden, and which conditions would make the stronger AI-native thesis durable.

The position

AI is already entering RAN optimization and network operations. The September 2026 question is no longer whether AI appears in the network, but where standards begin to make learned behavior **portable, observable, testable and governable** across implementations. The record supports a network that can become more AI-native without first converging on one universal accelerator or one clean-sheet architecture.

AI nativeness is a spectrum, not a binary label

For a non-AI reader: the ordering below asks how far learned behavior moves from an internal optimization toward an exposed lifecycle object and then toward authorized network action. Darker orange means a stronger AI-native assumption; these are analytical levels, not formal 3GPP categories.

AI-assisted - learned optimization stays inside a deterministic envelope. Examples: scheduler/link adaptation, traffic prediction, energy optimization; proof is persistent KPI gain and acceptable power cost.

AI-aware - model/data state, applicability, telemetry, feedback and fallback become visible to interfaces and assurance. Examples: RAN5 applicability/PICS/test, RAN3 AI/ML framework, SA6 lifecycle/inference data; proof is portability and repeatable test.

More deeply AI-native - agentic or intent-driven functions translate goals into network action under identity, authorization, policy, audit and rollback. Examples: CT3 AI-protocol study, OAM agents/intent and network-assisted Physical AI; proof is bounded autonomy plus safe operational control.

Historical context: continuity before discontinuity

Mobile generations rarely arrive as clean breaks. 4G and 5G progressively increased programmability, cloud integration and service differentiation; the emerging 6G record adds data, sensing, learned behavior and AI-aware control as ordinary engineering concerns. **AI-native does not mean AI everywhere.** It means model state, data provenance, inference conditions and lifecycle can become consequential enough that interfaces, operations and assurance must recognize them.

Commercial reality: three investment theses, not one

AI-for-RAN can justify itself through radio or operating KPIs. **AI-and-RAN** must additionally prove isolation, utilization, power and site economics for shared compute. **AI-on-RAN** adds a customer-demand test: repeat workloads, SLA terms, billing and margin. A live demo can advance one thesis without proving the others.

Why the September record matters

The important change since May is not that the industry discovered AI-RAN. Separate strands are acquiring institutional homes: radio groups narrow structural and AI/ML questions; SA addresses architecture, management and application lifecycle; CT has a dedicated AI-protocol study; assurance work is learning to express applicability, configuration and repeatability. The map of where future interoperability obligations could settle is becoming clearer.

The boundary that matters

In 5G, much learned behavior could remain an implementation detail. The stronger 6G question is whether model/data state, applicability, feedback, fallback and authorization become visible enough that another implementation can be trusted, compared or substituted. That is the point at which AI changes interoperability rather than simply improving an algorithm.

What 6G adds over the 5G baseline

The architectural novelty can shift upward. Familiar CU-DU and CP-UP boundaries can remain while data, model state, intent, sensing and assurance become ordinary objects of system design. For non-3GPP readers, this is the central distinction: the network need not look radically different in boxes and arrows for its interoperability burden to change materially.

Where value can appear first

The most credible near-term pattern is selective. Dense or high-traffic cells can justify AI-for-RAN where capacity or energy gains are measurable; industrial and private-network sites can justify richer local compute where sovereignty, locality or resilience matters; broader AI-on-RAN economics still depend on repeat external demand. Selectivity can be a rational design choice rather than evidence that AI-RAN failed to scale.

What not to assume

A more AI-native network does not imply a GPU at every site, a standardized foundation model, or autonomous control everywhere. The common denominator is narrower: learned state becomes important enough that another function, vendor, test system or operator needs a dependable way to observe its applicability, evidence, version and fallback. The report therefore separates AI nativeness from any one compute architecture or monetization story.

Source discipline

SRC primary record; **CNT** counted comparison; **INT** Apex synthesis with a stated condition/falsifier; **CTX** external context. Confidence is established / well supported / conditional. Page 2 carries the full vocabulary.

STANDARDS CLOCK

studies -> agreements -> specifications

What is established - and what remains open

The record supports three relatively firm statements: the 6G RAN higher-layer structural path is evolutionary rather than wholly clean-sheet; AI/ML is expanding from individual algorithms into data, lifecycle, management and protocol concerns; and implementation paths remain commercially plural. It does **not** yet establish a universal compute substrate, a complete 6G AI conformance regime, or a repeatable operator business for external AI workloads.

What materially changed after the May baseline

C20004008 [1] treated NVIDIA-Nokia as a market signal and deliberately left cost, energy, latency and multi-vendor proof unresolved. Singapore and Madrid do not close those questions, but they narrow where durable standards work is forming: the 6G RAN higher-layer structural baseline, AI/ML frameworks, data and lifecycle handling, management agents, protocol studies and assurance machinery.

Key insight 1 - structure can remain evolutionary while intelligence becomes more native

Madrid RAN3 retained the 5G HLS as the baseline for the 6G RAN higher-layer structure while leaving optimizations open where protocol distribution is unchanged. That constrains the clean-sheet narrative and shifts attention toward what the structure carries - data, model state, intent, sensing, service awareness and evidence.

Key insight 2 - intelligence contracts harden faster than the intelligence computer

The cross-record pattern is increasingly about contracts around learned behavior: information exchange, model or capability state, telemetry, authorization, fallback and proof. 3GPP can make those surfaces consequential without selecting GPU, CPU, NPU or ASIC as the universal substrate. Standards convergence and compute-platform convergence remain separate questions.

Key insight 3 - verification is moving toward architecture

A learned implementation carries state in the model version, dataset lineage, adaptation history and inference environment. RAN5 is still predominantly current-generation AI/ML machinery rather than a completed 6G regime, but it shows an institutional path toward applicability, PICS, configuration, data and repeatability artifacts.

Key insight 4 - lock-in can migrate upward

Open radio interfaces do not automatically create an open AI lifecycle. Switching friction can move into feature pipelines, model formats, quantization, runtime libraries, telemetry schemas, orchestration policy or agent frameworks. Lifecycle portability may become as important to procurement as accelerator choice.

Key insight 5 - heterogeneous deployment may be the equilibrium

The market already supports more than one credible path: AI functions on installed RAN compute, richer accelerator-centric infrastructure at selected sites, and external edge AI where demand justifies it. A mixed topology can be a stable economic outcome rather than a temporary phase.

Key insight 6 - standards may harden the seams before the market selects the machinery

Rel-21 decisions can make interface boundaries, data semantics, lifecycle assumptions and test obligations difficult to reverse while operators still compare hardware and runtime economics. This is not 3GPP choosing a GPU or model format; it is path dependence at the seams between components.

The deeper pattern

Taken together, the June-to-September evidence points toward a system that can remain **structurally familiar while becoming operationally more intelligent**. The durable change may be less visible than a new architecture box: more model/data state must be described, more evidence carried across organizational boundaries, and more of the AI lifecycle made observable enough to operate safely in a multi-vendor network.

What would strengthen the thesis next

The strongest next evidence would be reusable lifecycle semantics across more than one feature, multi-vendor assurance that survives model/runtime differences, and operator proof that gains persist after updates and across representative traffic. These would move AI-native from a feature narrative toward an interoperability property.

Industry paths are diverging by design

Nokia/NVIDIA represents the richer common-infrastructure thesis; Ericsson and Samsung demonstrate that meaningful AI-native behavior can remain credible on heterogeneous or existing RAN compute; SoftBank is testing a portfolio across common infrastructure, orchestration, live-network AI-in-RAN and Physical AI. The useful signal is not which camp wins first, but whether performance, power, portability and lifecycle proof survive materially different stacks.

Assurance is also switching-cost control

If only the incumbent can reproduce learned behavior because critical data, state or evaluation assumptions are hidden, switching cost rises even when network interfaces are formally open. Shared assurance semantics can preserve proprietary models while making the evidence required to trust, compare and replace them more portable.

Procurement consequence

For operators and buyers, the unit of comparison is shifting from the feature name toward the lifecycle evidence behind it: which dataset conditions were assumed, which model/runtime produced the result, how an update is validated, and whether fallback is observable. A standardized interface can still carry high switching cost if the proof needed to replace the learned implementation cannot travel with it.

EVIDENCE CLOCK

dataset -> model -> repeatability -> conformance

Where this is heading - conditionally

First layer - structurally evolutionary, operationally more AI-native. If Rel-21 preserves the current 6G RAN HLS baseline while selected AI/data/agent studies become stable work-item scope, novelty is more likely to concentrate in lifecycle, observability and control than wholesale protocol redistribution. This weakens if later Rel-21 decisions materially reopen that baseline.

First layer - assurance becomes a procurement surface. If model/data state becomes traceable enough for reusable applicability, test and fallback assumptions, buyers can compare learned implementations without seeing every proprietary detail. The contrary signal is proof tied to narrow features or vendor-specific runtimes.

Second layer - portable lifecycle. If common model-state, version, feedback and fallback semantics become implementation-independent, substitution can extend beyond interface compatibility into the AI lifecycle. If those semantics remain proprietary, nominally open interfaces may coexist with high switching cost.

Second layer - selective accelerated infrastructure. If multi-site utilization, power, cooling, isolation and TCO are favorable for some site classes, richer acceleration can expand in dense urban, regional-edge, enterprise and industrial locations while broad coverage still uses embedded or software AI.

Second layer - AI-on-RAN becomes a service only with repeat demand. Named paying customers, repeat workloads, SLA terms, billing and margin would convert distributed AI from an ecosystem capability into an operator business. If evidence remains project-specific, the nearer-term center of gravity stays with AI-for-RAN efficiency and automation.

Recent examples across the spectrum

AI-assisted: Ericsson/T-Mobile and SoftBank/Ericsson live-network scheduling/link-adaptation evidence; broad network optimization, capacity and energy use cases. **AI-aware:** RAN5 applicability/PICS/test, RAN3 framework work and SA6 lifecycle/inference-data handling; private networks and industrial edge are natural proving grounds because locality, sovereignty and assurance matter. **More deeply AI-native:** CT3 AI-protocol work, OAM agents/intent and Physical-AI experiments; robotics, V2X, drones/low-altitude systems and NTN become relevant where network state changes application outcomes.

The 2027-2029 horizon

Standards and commercial clocks will not close together. March 2027 package and Stage-1 decisions show which Rel-20 studies are mature enough for Rel-21 scope; Stage-2 and Stage-3 freezes follow in 2028, with coding/API freeze in 2029. A standards freeze can reduce technical uncertainty without proving site TCO, and commercial deployment can precede universal standardization if the needed interfaces are stable.

What would materially change this reading

Re-open the analysis if Rel-21 materially redistributes 6G RAN higher layers, AI lifecycle semantics remain inseparable from proprietary runtimes, assurance fails to generalize beyond narrow features, or multi-site economics show one compute model decisively superior. Reusable lifecycle interfaces, repeatable cross-vendor proof and named recurring AI-service customers would strengthen the more ambitious interpretation.

The strategic implication

The next high-value evidence will not be another AI-native label. It will be a durable interface, lifecycle rule, test procedure, multi-site operating result or repeat customer contract that changes what operators can safely substitute, scale or monetize.

Vertical lens: where higher nativeness has a reason to exist

Manufacturing / robotics can reward locality, deterministic behavior and network-aware sensing. **Automotive, V2X and drones** add mobility, positioning and safety-sensitive context. **Ports, utilities and private networks** can justify richer local compute through sovereignty and resilience. **NTN and constrained links** reward selective reduction, onboard inference and AI-assisted scheduling even when inference stays on the device or satellite. The common test is whether network state materially changes the application outcome.

Commercial proof is a ladder, not a label

AI-for-RAN: persistent KPI and power evidence. **AI-and-RAN:** real utilization, isolation and site TCO. **AI-on-RAN:** named repeat customers, SLA/billing and margin. **Agentic operation:** bounded authority, audit and safe rollback. Each closes a different risk, so one successful PoC cannot be used as a maturity score for the whole category.

What the next cycle can actually settle

Rel-21 can narrow interface, lifecycle, test and authorization assumptions without deciding the commercial winner among compute paths. Conversely, operator deployment can validate economics before every lifecycle detail is standardized. The useful watchpoint is therefore where a requirement becomes reusable across features and vendors; that is stronger evidence of AI-native system behavior than another isolated performance result.

Who reads what

AI / RAN experts: focus on lifecycle, assurance and interfaces. **Non-AI readers:** use the orange ladder as the mental model. **Operators / product:** compare proof burdens and site economics. **Policy / procurement:** watch portability, control and switching cost. Page 2 gives the full routes.

Decision questions

Standards: which assumption is becoming hard to reverse?
Operators: separate radio-performance value from external-AI revenue.
Policy/procurement: who controls the model, data, update path and operational override?
Product: what model/data/telemetry state must remain portable?

MARKET CLOCK

PoC -> field/live -> deployment -> repeat revenue

CONTENTS, EXHIBIT INDEX AND READER ROUTES

Thirteen parts, numbered part.section. Every numbered subsection below is generated from the built report and carries its resolved page number; the contents therefore remains synchronized when prose reflows.

Contents

CONTENTS, EXHIBIT INDEX AND READER ROUTES . . 2

1 AI-NATIVE 6G: HISTORICAL, PHILOSOPHICAL AND COMMERCIAL REALITY CHECK . . . 3
1.1 From transport to adaptive systems . . . 3
1.2 The hybrid operating model: deterministic baselines, AI-assisted optimization and explicit fallback . . . 3
1.3 Five tests of an AI-native claim . . . 3
1.4 The 5G lesson: capability is not monetization . . . 3

2 FROM APEX STANDARDS C20004008 TO SEPTEMBER 2026 . . . 3
2.1 The May baseline: two clocks and unresolved commercial proof . . . 3
2.2 Watch signals that materially strengthened . . . 4
2.3 Watch signals that remain unresolved . . . 4
2.4 How to read the longitudinal delta . . . 4

3 SINGAPORE TO MADRID: WHAT HARDENED FROM ONE PLENARY TO THE NEXT . . . 4
3.1 Singapore RAN#112: closing a framing stage . . . 4
3.2 The working-group interval: architecture, QoS, lifecycle and protocol questions become concrete . . . 4
3.3 Madrid RAN#113: the higher-layer structural baseline narrows . . . 4
3.4 What did not harden . . . 5

4 CROSS-TSG CONVERGENCE: DIFFERENT BODIES, ONE LEARNED SYSTEM . . . 5
4.1 RAN: learned behavior must survive radio timing and mobility . . . 5
4.2 RAN5: assurance turns learned behavior into a telecom product obligation . . . 5
4.3 SA2, SA5 and SA6: architecture, lifecycle and service exposure . . . 5
4.4 CT3 and security: protocol surfaces and authority . . . 5
4.5 Cross-TSG synthesis: the durable control point is the shared contract . 5

5 6G ARCHITECTURE, DATA AND MODEL LIFECYCLE 5
5.1 Evolutionary structure and expanding native intelligence . . . 5
5.2 Data becomes infrastructure . . . 6
5.3 Model state and lifecycle as operational objects . . . 6
5.4 Digital twins: from planning tool to repeatable evidence environment . 6
5.5 The architectural undercurrent: portability can fail above an open interface . . . 6

6 AI TRAFFIC, AGENTS, QOS AND PROTOCOL TRAJECTORY . . . 6
6.1 AI-aware traffic: the network may need richer abstractions . . . 6
6.2 Semantic communication: useful direction, difficult boundary . . . 7
6.3 Agent-to-agent communication: from vocabulary to protocol study . . 7
6.4 Intent-driven operation and the problem of translation . . . 7
6.5 Protocol outlook: standardize the seams, not necessarily the intelligence 7

7 VERIFICATION: FROM TEST EVENT TO LIFECYCLE ASSURANCE . . . 7
7.1 Why conventional RAN verification is not enough by itself . . . 7
7.2 The RAN5 precedent and its limits . . . 7
7.3 Generalization, drift and repeatability . . . 7
7.4 Digital twin, lab, OTA and field evidence serve different purposes . . 8
7.5 Verification as interoperability and switching-cost control . . . 8

8 INDUSTRY MOMENTUM AND COMPETING COMPUTE PATHS . . . 8
8.1 Ecosystem scale is no longer the main question . . . 8
8.2 Nokia/NVIDIA: richer common infrastructure and platform economics 8
8.3 Ericsson and Samsung: AI-native behavior without one mandatory accelerator . . . 8
8.4 SoftBank: operator portfolio rather than single-path architecture . . . 8
8.5 Competitive boundary: platform convergence is not implied by standards convergence . . . 8

9 AI-FOR-RAN, AI-AND-RAN AND AI-ON-RAN: THREE COMMERCIAL THESES . . . 8
9.1 AI-for-RAN: the shortest path from technical gain to operator value . 8
9.2 AI-and-RAN: shared infrastructure is an economic thesis . . . 9
9.3 AI-on-RAN: the network as a distributed AI service platform . . . 9
9.4 Why the three theses should not be scored as one market . . . 9
9.5 The 5G lesson revisited: repeatability closes more than novelty . . . 9

10 OPERATOR EVIDENCE, VERTICALS, ISAC AND PHYSICAL AI . . . 9
10.1 Operators as the reality check on vendor architecture . . . 9
10.2 Manufacturing and robotics: where locality and control can justify edge intelligence . . . 9
10.3 Video analytics, smart city and public-space sensing . . . 10
10.4 Automotive, drones and low-altitude systems: sensing and mobility converge . . . 10
10.5 Enterprise, ports, utilities and remote systems: selective deployment economics . . . 10
10.6 ISAC as a multiplier, not a guaranteed product . . . 10

11 COMMERCIAL READINESS AND OPERATOR ECONOMICS . . . 10
11.1 Performance is necessary; economics is decisive . . . 10
11.2 AI-for-RAN readiness: strongest evidence, still context-dependent . 10
11.3 Common accelerated infrastructure: utilization is the key variable . 10
11.4 External AI revenue: capability is not demand . . . 11
11.5 Procurement and policy: control surfaces become economic variables 11

12 FINDINGS, CONDITIONAL OUTLOOK AND WATCHLIST . . . 11
12.1 Finding: AI-native 6G is more radical in information than in structural split . . . 11
12.2 Finding: the institutional footprint of AI is expanding faster than any one algorithm . . . 11
12.3 Finding: verification is becoming part of product architecture . . . 11
12.4 Finding: platform lock-in may migrate upward . . . 11
12.5 Finding: heterogeneous deployment is a plausible equilibrium . . . 11
12.6 Layer 1 outlook: what follows if current consensus continues to harden.
12.7 Layer 2 outlook: downstream possibilities that require an additional condition . . . 12
12.8 2026-Q4 to 2029 watchlist: evidence that would materially change the reading . . . 12

13 METHOD, EVIDENCE BOUNDARY AND CLOSING CLAIMS . . . 12
13.1 Evidence hierarchy and status language . . . 12
13.2 What is not normalized . . . 12
13.3 Re-derivation triggers . . . 12
13.4 What this report does not claim . . . 12
13.5 Decision checklist . . . 13

APPENDIX A REFERENCE LEDGER - PRIMARY RECORDS FIRST . . . 14

Index of figures

Figure 1. Three clocks that should not be collapsed . . . 4
Figure 2. Evolutionary structure with an expanding native-intelligence layer . . . 5
Figure 3. Three layers that should not be collapsed . . . 6
Figure 4. Verification moves from a test event to a lifecycle . . . 7
Figure 5. Commercial proof is a ladder, not a label . . . 9

How to use this brief

The report is prose-first. Figures are retained only where they show a relationship more efficiently than another paragraph. Standards and architecture readers can begin with Parts 2-6; operator and product readers with Parts 4, 7-10; commercial and policy readers with Parts 7-12. The closing method section explains how a changed record triggers re-derivation.

Evidence-status rule

A contribution, study, work item, agreement, change request and normative specification are not interchangeable evidence states. The report uses the narrowest status supported by the cited record. Vendor/operator KPI results remain attributed and are not normalized into cross-vendor rankings because test conditions differ.

Conditional rule

Forward-looking passages use observation -> condition -> consequence -> falsifier. A first-layer outlook follows closely from current consensus. A second-layer outlook requires an additional technical, commercial or policy condition. No probability or winner score is assigned.

Evidence cutoff

18 September 2026. Madrid TSG#113 had just closed. Portal metadata can continue to update after the meeting; the edition therefore re-derives affected conclusions when a later primary record changes status, rather than preserving an older label for continuity.

Source discipline in detail

The page-one key is intentionally compact. The working rule used in the body is stricter: every paragraph is classified by what kind of statement it makes, not by how persuasive the conclusion sounds.

SRC **Source record.** Use when the statement is directly supported by a named primary record - for example a 3GPP portal status, meeting agreement, Technical Report, Work Item, Change Request or an ITU/O-RAN primary record. Status language follows the record itself.

CNT **Counted comparison.** Use for a count, delta or longitudinal comparison produced by a stated rule. Counts are descriptive and are not converted into influence scores, market-share measures or implied votes.

INT **Interpretation.** Use for Apex Standards synthesis across records. Interpretations identify the transition from record to implication; where the conclusion depends on a future fact, the text states the condition and the observation that would weaken it.

CTX **External context.** Use for vendor/operator releases, alliance activity, trade press and cross-SDO material outside the formal 3GPP decision record. Vendor KPI claims remain attributed and are not silently normalized against other trials.

Confidence vocabulary

Established is reserved for a direct, stable record or a narrow conclusion with little interpretive distance. **Well supported** indicates a synthesis backed by multiple consistent records but still requiring interpretation. **Conditional** identifies an outlook that requires an additional technical, commercial or policy condition. These labels describe evidentiary footing, not importance or probability.

Status ladder

Proposal -> discussion -> agreement -> study -> Work Item -> specification text -> conformance / implementation evidence -> deployment -> repeat commercial use. Real programs do not always move linearly through this sequence, but the ladder is useful because it prevents a contribution from being described as consensus and prevents a PoC from being described as a market.

Reader routes

For a fast executive read: page 1, Parts 2-3, Part 8, Part 12. For standards strategy: Parts 3-7 and the reference ledger. For product and procurement: Parts 5, 7-9, 11-13. For operator or vertical strategy: Parts 8-12. The sections are intentionally connected by prose so a reader can also follow the document linearly without reconstructing logic from tables.

Source hygiene

When a primary record and an index page disagree on a date, this edition uses the date printed on the individual record being cited. Combined source entries are split when one hyperlink cannot support both claims. The prior Apex Standards C20004008 report is used only as the May analytical baseline; it is not treated as a standards authority.

Recurring terms used narrowly

AI-native means that learned behavior, its state and its lifecycle become consequential system variables; it does not mean every network function is learned. **AI-for-RAN** means AI used to improve RAN performance or operations. **AI-and-RAN** refers to shared or coordinated infrastructure for RAN and AI workloads. **AI-on-RAN** refers to application/AI workloads executed on or near distributed telecom infrastructure. **HLS baseline** refers here to the 6G RAN higher-layer structural split discussed by RAN3, not to the entire end-to-end 6G architecture.

What to look for in later editions

The highest-value changes are those that alter a control surface: Rel-21 scope that changes protocol distribution; stable data/model lifecycle semantics; reusable assurance assumptions; multi-vendor reproducibility; multi-site power/TCO evidence; repeat external-AI customers; or security/governance rules that materially change agent authority. Announcements that do not move one of these boundaries remain context rather than a changed conclusion.

The contents therefore doubles as an audit map: the reader can move from an executive statement to the standards transition, the implementation evidence, the commercial boundary and finally the primary record without treating any one layer as a substitute for the others.

1 AI-NATIVE 6G: HISTORICAL, PHILOSOPHICAL AND COMMERCIAL REALITY CHECK

1.1 From transport to adaptive systems

INT WELL SUPPORTED - The useful historical progression is not a simple list of peak data rates. Earlier generations primarily expanded what information could be carried and how flexibly services could be delivered. 4G made all-IP broadband the default; 5G added cloudification, service differentiation and more explicit programmability. The 6G discussion adds another dimension: the network may increasingly observe, infer, predict, sense, optimize and coordinate computation as part of normal operation. That is a systems change even if familiar structural interfaces survive.

INT WELL SUPPORTED - This distinction matters because the word native is often used as if it required a clean architectural break. In deployed telecommunications, native capability more often emerges through the accumulation of interfaces, procedures and operational obligations that become too important to remain proprietary side logic. AI becomes meaningfully native when data, model state, inference conditions, update behavior and fallback affect interoperable system behavior - not merely when a vendor includes a neural model inside a product.

INT WELL SUPPORTED - Every mobile generation arrives with two stories. The first is technical: what the network can do. The second is economic: what those capabilities will be worth. Those stories rarely mature at the same speed. 5G made the network more programmable, cloud-aware and service-differentiated, but it also reminded the industry that technically successful capabilities do not automatically produce equally transformative new operator revenue. That historical separation is the right place to begin an AI-native 6G report.

INT WELL SUPPORTED - The philosophical shift is deeper than adding machine learning to a scheduler. A conventional network function can be described largely as protocol plus software plus hardware. A learned function adds a model, training data, an inference environment, update rules, performance feedback and a lifecycle. Once those become part of the engineering object, the standards question changes from “what does this function do?” to “what remains interoperable and verifiable when behavior depends partly on data and learned state?”

SRC ESTABLISHED - That change expands the network's design vocabulary. The system is no longer asked only to transport information. It may increasingly observe, infer, predict, optimize, sense, coordinate compute and expose context to applications. ITU's IMT-2030 framework already places AI and Communication and ISAC inside the formal 6G usage-scenario landscape, while 3GPP Release 20 combines continuing 5G-Advanced work with studies that will feed Release 21 normative work. [11],[12]

1.2 The hybrid operating model: deterministic baselines, AI-assisted optimization and explicit fallback

INT WELL SUPPORTED - Carrier networks have deadlines, state machines, emergency procedures, lawful obligations and operational accountability that do not disappear because a learned function performs well on average. The near-term engineering pattern is therefore likely to remain hybrid. Learned functions can optimize scheduling, channel estimation, prediction, energy or mobility where they produce measurable gain, while deterministic control bounds failure modes and provides a safe path when model confidence, context or data quality deteriorates.

INT WELL SUPPORTED - Hybrid does not mean temporary or inferior. In safety- and availability-sensitive systems it can be the stable architecture: conventional logic specifies the envelope; learned behavior searches inside it; monitoring decides when the learned path is applicable; and fallback preserves continuity when assumptions fail. That architecture also produces a clearer standards problem because interfaces can describe applicability, state and fallback without forcing every vendor to reveal or standardize the internal model.

INT WELL SUPPORTED - Generational labels make sharp discontinuities easy to draw on slides. Deployed networks usually evolve through coexistence, migration, backward-compatible economics and selective replacement. That matters because the public language around 6G is already full of “native” claims. The stronger interpretation is not that 6G discards 5G architecture, but that more functions may become data-aware, model-aware and context-aware while the network preserves interfaces and deployment paths operators already know how to run.

INT WELL SUPPORTED - Technical test. Does the learned behavior improve a network or service KPI under realistic conditions, not only in a favored dataset or lab topology? AI-for-RAN scheduling and link adaptation currently have the clearest public answer because live-network results can be expressed in spectral efficiency, throughput or capacity. [15]-[17]

INT WELL SUPPORTED - Verification test. Can the result be reproduced across datasets, environments, versions and implementations? AI introduces failure modes that are awkward for conventional pass/fail thinking: distribution shift, model drift, hidden dataset dependence, retraining effects and changing inference hardware. The test system therefore has to describe not only the RF condition but the model state and data assumptions under which the result is valid. [5],[22]

1.3 Five tests of an AI-native claim

INT WELL SUPPORTED - A technical claim should first show measurable benefit under realistic radio or service conditions. A verification claim should then show that the result survives changes in dataset, environment, model version and implementation. A standardization claim asks whether independent systems can exchange enough information to interoperate without assuming a hidden vendor stack. A deployment claim asks whether operators can run the feature safely at scale. An economic claim asks whether the capacity, experience, automation or revenue benefit exceeds the additional compute, power and lifecycle cost.

INT WELL SUPPORTED - These tests deliberately separate strong algorithmic evidence from strong product evidence. A model can outperform a baseline and still fail to become a telecom feature if it is difficult to reproduce, expensive to operate, impossible to validate after updates or tightly coupled to one environment. Conversely, a modest model improvement can become strategically important if it is repeatable, portable and cheap enough to deploy across a large installed base.

INT WELL SUPPORTED - Standardization test. Can independently implemented systems interoperate without assuming one vendor's model runtime or closed data path? 2026 work is increasingly concentrated on interfaces, lifecycle, management, data collection and protocol surfaces precisely because the standard does not need to prescribe one model to make learned behavior interoperable. [4]-[10]

INT WELL SUPPORTED - Deployment test. Can an operator run the capability safely and continuously at scale? A field trial proves more than a lab demo, but it still may not prove fleet-wide energy, resilience, upgrade behavior, observability or site economics. That is why the proof ladder in this report runs from concept to lab to field to live network to repeatable deployment, with repeatable revenue as a separate final step.

INT WELL SUPPORTED - Economics test. Does the performance, capacity, automation or new revenue exceed added compute, power and operational cost? AI-for-RAN often has a direct value chain - KPI gain to capacity or quality to cost deferral. AI-and-RAN requires an additional thesis: shared accelerated infrastructure must preserve carrier-grade RAN behavior while raising utilization. AI-on-RAN is harder still because someone must pay for distributed AI capability.

INT WELL SUPPORTED - These five tests prevent the report from confusing capability with consequence. They also explain why a standards agreement can be strategically important without immediately changing the market, and why a market partnership can move headlines much faster than conformance, interoperability and economics can move.

1.4 The 5G lesson: capability is not monetization

INT WELL SUPPORTED - 5G is a useful commercial reality check because it demonstrated that technical capability and incremental revenue are different clocks. A network generation can improve capacity, latency, flexibility and coverage without automatically creating a new operator business model of equal magnitude. AI-RAN should be read with that history in mind. AI-for-RAN may close its value chain through internal network economics, while AI-and-RAN and AI-on-RAN depend on infrastructure-utilization and customer-demand assumptions that are not resolved by technical performance alone.

INT WELL SUPPORTED - The consequence for this report is methodological: a standards agreement is not treated as a business case, and a commercial partnership is not treated as a standards outcome. The analysis becomes more useful when it asks what each evidence class can legitimately close. Standards can reduce interoperability uncertainty; field trials can reduce performance uncertainty; multi-site operation can reduce scaling uncertainty; customer contracts can reduce demand uncertainty. None substitutes for all the others.

Reality check

The most important 6G discontinuity may be in what the network is expected to observe, expose and govern, while much of the structural deployment path remains evolutionary. That is why this report tracks architecture, lifecycle, verification and economics separately rather than using AI-native as a single maturity label.

2 FROM APEX STANDARDS C20004008 TO SEPTEMBER 2026

2.1 The May baseline: two clocks and unresolved commercial proof

SRC ESTABLISHED - The 24 May 2026 **Apex Standards C20004008** brief [1] framed the NVIDIA-Nokia relationship as a market signal whose durable meaning would be decided later by patents, standards and deployment proof. It placed Nokia nearer to telecom control and AI-native network procedures and NVIDIA nearer to accelerated compute, CUDA, memory movement, power and thermal scaling. Crucially, it did not equate strategic alignment with operator economics.

INT WELL SUPPORTED - That earlier discipline matters because the strongest September conclusions can now be expressed as deltas rather than hindsight. The question is not whether the May report was bullish or cautious; it is which stated watch signals acquired a more formal home, which remained open, and which moved in a different direction from the original market narrative. This longitudinal method makes the update more informative than a new snapshot.

SRC **ESTABLISHED** - Apex Standards C20004008, published on 24 May 2026, framed the NVIDIA-Nokia partnership as a signal sitting between two clocks: market activity could move immediately, while patents, standards and deployment proof would speak later. It placed Nokia closer to telecom control and AI-native network procedures, NVIDIA closer to compute infrastructure and CUDA-centered execution, and warned that the commercial case still depended on cost, energy, latency and multi-vendor proof. [1]

SRC **ESTABLISHED** - Four months later, the direction of several watch items is clearer. Agentic AI is no longer visible only through scattered RAN3, SA2 and SA5 contributions; CT3 now has a PCG-approved Rel-20 study and TR 29.832 as a dedicated protocol home. NG-RAN AI/ML Phase 3 has progressed through both June and September checkpoints. RAN5 has moved further into AI/ML conformance and data-collection procedures. SA6 has approved additional application-enablement changes covering inference data, model maintenance and performance feedback. [4]-[10]

INT **WELL SUPPORTED** - That does not mean every May hypothesis hardened equally. Token and AI-aware QoS moved further into SA2 architecture treatment, but the final normative model remains open. Semantic communication remains a useful family of ideas spanning service and physical-layer discussions, yet the record still does not support treating “semantic communication” as one adopted end-to-end 3GPP stack. GPU/CUDA adoption has gained commercial momentum, but no 3GPP agreement selects a universal accelerator architecture. [2],[9],[14]-[17]

2.2 Watch signals that materially strengthened

SRC **WELL SUPPORTED** - The formal footprint of AI-related work widened. NG-RAN AI/ML Phase 3 continued to progress; CT3 established TR 29.832 as a dedicated Rel-20 draft study for AI protocol; SA5 advanced 6G management and orchestration; SA2 continued the 6G architecture study; and RAN5 broadened current-generation AI/ML applicability, PICS, test and data-collection work. [4]-[10],[26],[28]

INT **WELL SUPPORTED** - Taken together, these movements strengthen the proposition that learned behavior is becoming a systems concern rather than a feature isolated in one RAN working group. The key change is institutional: more of the telecom organization now has a defined reason to care about model or data behavior, service context, authorization, lifecycle, testability or protocol exposure.

INT **WELL SUPPORTED** - The largest change is therefore not a single technology choice. It is the widening of the control surface. AI now touches radio behavior, data collection, model lifecycle, management, protocol exposure, conformance and applications. The standardization question increasingly becomes how to keep those surfaces observable, portable and auditable when implementation below them remains competitive.

INT **WELL SUPPORTED** - C20004008 also warned that lock-in could come from libraries, agent frameworks, data descriptions, observability and closed management loops rather than from radio hardware alone. The September record makes that caution more relevant, not less. As standards reach upward into lifecycle and agent behavior, proprietary coupling can migrate into model formats, telemetry schemas, orchestration policy or runtime assumptions even when the radio interface itself remains open. [1],[6],[7],[10]

INT **WELL SUPPORTED** - The retest therefore changes the language of the watchlist. The May question was whether AI traffic, agents and model lifecycle would become visible across multiple 3GPP layers. The September question is where that visibility becomes a durable interface contract, which parts remain vendor implementation, and whether operator economics support the additional compute and operational machinery.

2.3 Watch signals that remain unresolved

INT **WELL SUPPORTED** - Compute-platform convergence remains unresolved. Nokia/NVIDIA momentum strengthened, but Ericsson and Samsung evidence supports alternative paths in which AI features run on existing or heterogeneous RAN compute. Likewise, external AI workloads and Physical-AI demonstrations broaden the commercial imagination without yet proving repeatable demand, site economics or customer willingness to pay at scale.

INT **WELL SUPPORTED** - The original lock-in concern also remains open in a more nuanced form. Even if radio interfaces stay standardized, portability can be constrained by feature pipelines, model formats, runtime libraries, data schemas, observability and orchestration policy. The September evidence therefore shifts the lock-in question from a simple accelerator dependency toward the entire AI lifecycle.

Figure 1. Three clocks that should not be collapsed

1	**STANDARDS CLOCK** requirements -> studies -> agreements -> specifications
2	**MARKET CLOCK** partnership -> PoC -> field trial -> deployment -> revenue
3	**EVIDENCE CLOCK** dataset -> model -> test -> repeatability -> conformance

Commercial significance rises only when standards maturity, market adoption and reproducible evidence become sufficiently aligned.

2.4 How to read the longitudinal delta

INT **WELL SUPPORTED** - The report treats strengthening, weakening and remaining-open as different outcomes. A signal strengthens when it gains an approved study, agreed baseline, stable interface discussion or reproducible field evidence. It weakens when a later record narrows or rejects the premise. It remains open when activity increases but the decisive evidence - for example multi-site TCO or repeat revenue - is still absent. This vocabulary prevents momentum from being mistaken for closure.

3 SINGAPORE TO MADRID: WHAT HARDENED FROM ONE PLENARY TO THE NEXT

3.1 Singapore RAN#112: closing a framing stage

SRC **ESTABLISHED** - TR 38.914, *Study on 6G Scenarios and requirements*, reached version 1.0.0 at RAN#112 and was moved under change control in June 2026. [3] That does not mean 6G requirements were permanently frozen, but it marks a change in work character: the industry can spend less time asking whether a broad scenario belongs in 6G and more time translating the accepted direction into radio, architecture, management and test questions.

SRC **ESTABLISHED** - The June cycle also clarified the Release 21 schedule. The explicit milestones - March 2027 package approval / Stage-1 freeze, June 2028 Stage-2 freeze, December 2028 Stage-3 functional freeze and March 2029 ASN.1/OpenAPI freeze - provide a calendar against which study maturity can be judged. [11] The schedule does not tell the industry which AI concepts will survive, but it tells contributors when ambiguity has to convert into scope.

SRC **ESTABLISHED** - Singapore matters because it closed one phase of 6G framing. RAN#112 completed the 6G scenarios and requirements study and placed TR 38.914 under change control. That moved the debate away from an open-ended catalogue of possible 6G use cases and toward the harder work of radio, architecture, management and test design. [3]

SRC **ESTABLISHED** - Madrid matters because some of those design questions now have recognizable baselines. The clearest example is the higher-layer split. RAN3 agreed to support CU-DU split and CU-CP/CU-UP separation with the 5G higher-layer split as the baseline, while still allowing optimizations that do not alter protocol distribution. This is a strong reality check against any narrative that assumes AI-native 6G requires a clean-sheet RAN structure. [2]

3.2 The working-group interval: architecture, QoS, lifecycle and protocol questions become concrete

INT **WELL SUPPORTED** - Between the plenaries, the important work was not a single headline vote. Working groups translated broad 6G scope into candidate information flows, use cases, architecture alternatives, protocol questions and assurance artifacts. SA2's 6G architecture work, RAN3's framework and data discussions, SA5's management studies and CT3's protocol study show how a generational vision becomes a set of engineering contracts that can later be accepted, narrowed or rejected.

INT **WELL SUPPORTED** - This interval is where contribution volume is easiest to misread. Many proposals appear because the option space is still open. A high document count therefore signals attention, not consensus. The stronger evidence is what survives into moderator summaries, approved studies, agreed baselines, adopted change requests or stable specification text. The report privileges those dispositions over raw contribution frequency.

SRC **ESTABLISHED** - At the same time, intelligence is becoming more native around that structure. NG-RAN AI/ML Phase 3 moved from 30 to 50 percent completion by the September checkpoint. RAN3 continues 6G AI/ML use-case and framework work. SA6 application enablement gained additional lifecycle and inference-data requirements. OAM studies advanced management-agent and intent concepts. CT3 continued the dedicated 6G AI protocol study. [4],[6],[7],[10]

SRC **ESTABLISHED** - The verification track also moved. Singapore approved AI/ML UE-conformance material for CSI-prediction applicability and PICS. By September, RAN5 had agreed further AI/ML test and data-collection changes and forwarded a package to RAN#113. That movement is strategically important because it converts learned behavior from an algorithm demonstration into telecom proof artifacts: applicability, configuration, data collection, expected behavior and eventually repeatable test cases. [5],[28]

3.3 Madrid RAN#113: the higher-layer structural baseline narrows

SRC **ESTABLISHED** - RAN3 reported that 6G RAN architecture shall support the higher-layer split, including CU-DU split and CP-UP separation, with the 5G HLS as the baseline; optimizations that do not affect protocol distribution are not precluded. [2] This wording should be read narrowly. It is a **6G RAN higher-layer structural baseline**, not an endorsement of every 5G architectural element as the complete 6G system design.

INT **WELL SUPPORTED** - The agreement matters because it reduces one class of migration risk. Vendors and operators can reason about new 6G intelligence, sensing, service and data functions without simultaneously assuming a wholesale redistribution of familiar higher-layer protocols. That favors evolutionary insertion points and allows innovation to concentrate in new semantics, state exposure and behavior around the existing split.

INT WELL SUPPORTED - ISAC shows the same pattern. The discussion is no longer only about whether sensing belongs in 6G. Follow-on RAN work is moving toward measurements, architecture and implementation constraints. The commercial implication is not that sensing revenue is proven, but that network-assisted perception is becoming technically concrete enough to require the same lifecycle of interfaces, privacy, calibration and validation that connectivity features already face. [2],[29]

3.4 What did not harden

INT WELL SUPPORTED - Madrid did not select a universal AI model, accelerator, runtime, model-exchange format or operator monetization model. Nor did current RAN5 AI/ML work become a finished 6G certification regime. These absences are analytically important because they define the remaining competitive surface. Standards can narrow how learned behavior interacts with the network while leaving implementation choice and business architecture open.

Singapore-to-Madrid reading
The delta is cumulative rather than theatrical: a controlled requirements baseline, a fixed Rel-21 calendar, a retained RAN HLS baseline, broader AI/ML responsibility, a dedicated AI-protocol study and more concrete assurance machinery. Together they narrow where future normative work can land without implying that the market has selected one compute path.

4 CROSS-TSG CONVERGENCE: DIFFERENT BODIES, ONE LEARNED SYSTEM

4.1 RAN: learned behavior must survive radio timing and mobility

INT WELL SUPPORTED - RAN is where AI claims encounter the discipline of radio systems. Scheduling, link adaptation, channel estimation, beam behavior, mobility and positioning are constrained by timing, measurement availability, signaling overhead and the need to coexist with legacy behavior. The strategic signal is therefore not merely that AI appears in a contribution; it is that use cases acquire stable information exchange, bounded procedures and eventually testable behavior.

SRC WELL SUPPORTED - RAN#113 records show continuing AI/ML work in both NG-RAN and air-interface domains, while RAN3 identifies AI/ML use cases and framework and the 6G data framework among its study topics. [2],[4],[27] The combination connects algorithmic questions to architecture and data handling rather than treating model performance as a self-contained research result.

INT WELL SUPPORTED - RAN is where learned behavior has to survive radio timing, mobility, scheduling and architecture constraints. The most important evidence is therefore not that "AI" appears in a contribution, but that a use case acquires stable information exchange, bounded procedures and eventually testable behavior. NG-RAN Phase 3 and AI/ML air-interface work are useful because they make model support part of normal RAN engineering rather than an external research attachment. [4],[27]

INT WELL SUPPORTED - RAN5 plays a different role. Its importance is institutional: it translates learned behavior into familiar telecom assurance objects. Applicability, PICS, test configuration, data collection and pass/fail expectations are the bridge from an algorithm claim to procurement-grade evidence. As AI features diversify, the question will be how much of the model state and dataset assumption must become explicit before two vendors can reproduce the same result. [5],[28]

4.2 RAN5: assurance turns learned behavior into a telecom product obligation

SRC WELL SUPPORTED - The strongest direct conformance evidence in this report is still primarily NR/current-generation AI/ML machinery: applicability, PICS, test configuration and data-collection artifacts around identified AI/ML features. [5],[28] It should not be described as an existing 6G conformance regime.

INT WELL SUPPORTED - Its strategic importance is procedural. RAN5 demonstrates how an algorithmic feature can be translated into artifacts familiar to procurement, certification and acceptance testing. That pathway is likely to matter more as learned functions become updateable and context-dependent because the assurance object must capture not only an input/output pair but also enough model, data and configuration state to make a result interpretable.

INT WELL SUPPORTED - SA broadens the problem from radio optimization to system architecture, service requirements, management and applications. SA2 is wrestling with AI traffic characteristics, packet importance, dynamic flows and architecture support. SA5 is examining management agents, observability, intent and policy. SA6 exposes model and inference lifecycle to applications. Together they make the system concern larger than "AI improves RAN performance." It becomes "how does the network operate learned functions over time?" [7],[9],[10]

4.3 SA2, SA5 and SA6: architecture, lifecycle and service exposure

SRC WELL SUPPORTED - SA2 owns the system-architecture view through TR 23.801-01; SA5 owns the management and orchestration view through TR 32.801-01; SA6 continues application-enablement work around inference and AI/ML lifecycle. [7],[10],[26] These groups do not duplicate RAN work. They expose the service, management, application and cross-domain obligations that arise once AI affects behavior outside a single radio algorithm.

INT WELL SUPPORTED - This is where model lifecycle becomes operational rather than academic. Operators need to know what is deployed, what version is active, what data or context it expects, how performance is observed, how intent is translated into action and how an update can be rolled back. A system can remain vendor-differentiated internally while still requiring common semantics at those control points.

INT WELL SUPPORTED - CT becomes relevant when agents, AI services or AI functions must communicate through standardized protocol surfaces. The existence of TR 29.832 is therefore more important than any single early proposal inside it. A dedicated study means that interoperability pressure has moved close enough to protocol definition that CT must decide what belongs inside 3GPP, what can reuse external mechanisms, and what security or authorization semantics are required. [6]

INT WELL SUPPORTED - The cross-TSG synthesis is that AI-native 6G may be defined less by one neural block than by contracts between layers: RAN timing and measurements, SA lifecycle and policy, CT protocols, and RAN5 proof. The durable control point is the interface shared across them. That is also where hidden proprietary assumptions are most costly, because they can make a nominally open architecture operationally difficult to substitute.

4.4 CT3 and security: protocol surfaces and authority

SRC WELL SUPPORTED - TR 29.832 gives AI protocol questions a dedicated CT3 study home, while TR 33.801-01 provides the broader 6G security study context. [6],[8] The important boundary is not whether every AI-agent protocol originates in 3GPP. It is which mobile-system-specific identity, authorization, discovery, policy or mobility semantics 3GPP chooses to define, profile or require.

INT WELL SUPPORTED - Agentic operation raises an authority question that ordinary application traffic does not. An agent that can observe, recommend or act on network state needs a chain of permission, policy, audit and rollback. The more autonomous the action, the more the protocol surface becomes inseparable from governance. That is why agent interoperability and agent authorization should be read together rather than as separate application conveniences.

INT WELL SUPPORTED - This is why the report treats cross-layer convergence as stronger evidence than repeated use of the same vocabulary. "Agent," "intent," "semantic," "model" and "data" can appear across groups while meaning different things. Convergence is demonstrated only when one layer can consume another layer's contract without inventing proprietary translation.

4.5 Cross-TSG synthesis: the durable control point is the shared contract

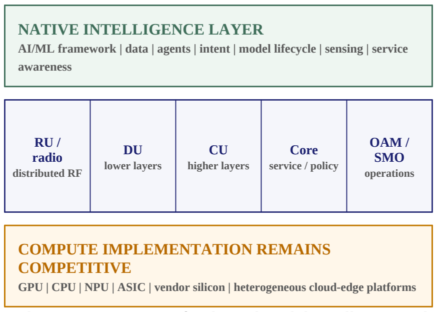
INT WELL SUPPORTED - The same learned system can therefore create obligations in four different places: RAN timing and information exchange, SA service and lifecycle semantics, CT protocol behavior, and RAN5 assurance. Cross-TSG convergence does not imply one monolithic AI architecture. It implies that hidden assumptions become harder to tolerate because a contract must survive organizational boundaries as well as vendor boundaries.

INT WELL SUPPORTED - This is one of the report's most important undercurrents. A model may remain proprietary, yet the interfaces that describe when it is active, what state it consumes, what capabilities it exposes, how it is monitored and what evidence validates it can become shared infrastructure. Those seams can create interoperability or lock-in long before the market selects a common accelerator.

5 6G ARCHITECTURE, DATA AND MODEL LIFECYCLE

5.1 Evolutionary structure and expanding native intelligence

Figure 2. Evolutionary structure with an expanding native-intelligence layer



Architecture can preserve familiar splits while intelligence and lifecycle obligations expand above them. Compute substrate choices remain implementation competition rather than 3GPP consensus.

INT WELL SUPPORTED - The architecture figure deliberately separates the structural RAN split from the intelligence layer above it. CU-DU and CP-UP continuity can coexist with new AI/ML framework functions, data handling, agent behavior, sensing, lifecycle and service-awareness obligations. That separation is more useful than a binary clean-sheet-versus-evolution debate because it shows where change can enter without forcing every layer to move at once.

INT WELL SUPPORTED - For deployment strategy, that architecture supports incrementalism. An operator can introduce AI-assisted functions through software and selective compute upgrades, centralize some functions where transport allows, and preserve distributed radio behavior where latency or resilience requires it. The result can be AI-native in operational semantics while remaining evolutionary in physical topology.

INT WELL SUPPORTED - The Madrid higher-layer-split agreement is valuable because it separates two debates that are often mixed together. Structural architecture can remain evolutionary while intelligence becomes more native. Retaining the 5G HLS baseline reduces migration uncertainty and preserves familiar CU-DU and control/user-plane separation, yet it does not prevent new data, model, sensing or agent functions from entering the system. [2]

INT WELL SUPPORTED - That is a more plausible deployment path than assuming every AI-native function must arrive with a new split. Operators can absorb new capabilities through software, interfaces and selective compute upgrades, while deferring major structural changes until the economics of transport, power and site design justify them. The result may be a 6G network that is intellectually more different than it is physically different.

INT WELL SUPPORTED - The architecture debate then moves upward. Where does model state live? How does the network know which model version produced a result? What telemetry is necessary to detect degradation? How is training or inference data collected, protected and versioned? Which agent is authorized to change network state? How does a UE, edge function or network service discover an AI capability? Those questions create new contracts even if the DU/CU split looks familiar. [6]-[10]

5.2 Data becomes infrastructure

INT WELL SUPPORTED - When performance depends on learned behavior, data stops being only an offline research input. Measurement provenance, feature construction, sampling conditions, labeling, privacy, retention and distribution shift can affect whether a model remains valid in operation. Those are lifecycle questions with direct consequences for troubleshooting and accountability because a performance change may originate in data conditions even when software and hardware are unchanged.

INT WELL SUPPORTED - This pushes the architecture toward explicit data contracts. Not every dataset or feature pipeline must be standardized, but the network may need common ways to describe availability, quality, timing, scope or permission so that distributed functions can reason about what evidence they are consuming. The 6G data-framework discussion is therefore strategically important even before exact interfaces are normative.

INT WELL SUPPORTED - Data becomes infrastructure once learned behavior enters a network function. Measurements that were previously transient telemetry can become training inputs, inference features or regression evidence. That raises questions of provenance, representativeness, retention, privacy and versioning. A model lifecycle that cannot explain what changed between two versions is difficult to govern even if the interface around the model is standardized.

INT WELL SUPPORTED - SA6 and SA5 activity is therefore strategically important. Application enablement now includes more explicit model maintenance, inference-data and performance-feedback concepts, while OAM studies develop management-agent and policy structures. These are not glamorous radio features, but they are the machinery that determines whether an AI function can be updated, observed and rolled back in production. [7],[10]

5.3 Model state and lifecycle as operational objects

INT WELL SUPPORTED - Traditional software lifecycle management already tracks versions, dependencies and rollback. AI adds additional state: model version, training lineage, calibration or adaptation state, feature expectations and performance feedback. If that state affects network behavior, an operator needs enough visibility to know what changed when a KPI shifts. Lifecycle transparency is therefore a reliability function, not only an MLOps convenience.

INT WELL SUPPORTED - Portability becomes a second question. A network interface can be open while model lifecycle remains proprietary. If one implementation describes model state, telemetry and fallback differently from another, substitution can require translation layers even when packets and procedures are standardized. Common lifecycle semantics can reduce that friction without forcing vendors to share model weights or training data.

SRC ESTABLISHED - Digital twins fit this lifecycle because they create a controlled environment between offline simulation and live network exposure. O-RAN identifies AI/ML training, evaluation and testing, network-test automation, planning, energy saving and site-specific optimization among the principal DT-RAN use cases. The key shift is from “digital twin as planning model” to “digital twin as repeatable evidence environment.” [23]

INT WELL SUPPORTED - That evolution also changes the economics of verification. If a digital twin can reproduce network states, generate rare conditions and replay them after a model update, the operator can reduce dependence on risky live experimentation. But the twin itself then needs validation: its channel, traffic, mobility and infrastructure models must be accurate enough that a good simulated result predicts field behavior.

5.4 Digital twins: from planning tool to repeatable evidence environment

CTX WELL SUPPORTED - O-RAN digital-twin work identifies AI/ML training, evaluation and testing, network-test automation, planning, energy saving and site-specific optimization as major use cases. [23] Keysight, Samsung and NVIDIA have separately demonstrated an end-to-end validation workflow linking scenario/data generation, model training and benchmarking. [22] These are ecosystem signals rather than 3GPP normative requirements, but they show why digital twins are becoming part of the AI lifecycle.

INT WELL SUPPORTED - A digital twin is valuable when it turns an expensive or unsafe field condition into a repeatable experiment. It can expose a model to mobility, interference, traffic or failure scenarios before deployment and preserve the context needed to compare versions. The limit is equally important: the twin itself must be validated. A highly repeatable simulation is not automatically predictive of field behavior if propagation, traffic, hardware or control assumptions are wrong.

INT WELL SUPPORTED - Model portability is the other major control point. Open network interfaces do not guarantee an open AI lifecycle. A model may still depend on proprietary feature pipelines, runtime libraries, quantization, accelerator kernels or telemetry semantics. Standards that expose lifecycle hooks without addressing those dependencies may produce nominal interoperability with practical switching friction.

INT WELL SUPPORTED - The watch signal is therefore the appearance of common schemas and version semantics across RAN, SA and management work. If operators can identify model version, required input, performance feedback and fallback behavior without depending on one vendor runtime, AI-native networking becomes materially more substitutable. If not, the model lifecycle becomes the new place where platform gravity accumulates.

5.5 The architectural undercurrent: portability can fail above an open interface

INT WELL SUPPORTED - The emerging AI-native control surface therefore sits above and beside classic protocol interoperability. A radio interface can be fully standardized while model formats, feature pipelines, runtime libraries, telemetry schemas and orchestration policy remain proprietary. The more performance and operational state moves into those layers, the more switching cost can migrate upward even as traditional interfaces remain open.

Figure 3. Three layers that should not be collapsed

STRUCTURE
HLS | CU-DU | CP-UP | protocol distribution

INTELLIGENCE CONTRACTS
data | model state | telemetry | intent | authorization | fallback | proof

IMPLEMENTATION
GPU | CPU | NPU | ASIC | vendor silicon | cloud-edge placement

Standards can narrow structural and intelligence contracts while implementation competition remains open. This distinction is central to the September 2026 reading.

6 AI TRAFFIC, AGENTS, QOS AND PROTOCOL TRAJECTORY

6.1 AI-aware traffic: the network may need richer abstractions

INT WELL SUPPORTED - AI applications can generate traffic patterns that are bursty, iterative, asymmetric and sensitive to importance or deadlines in ways that are not fully captured by simple throughput labels. Earlier 6G material discussed token characteristics and tolerance; later architecture work considers AI service behavior, packet importance and changing flow characteristics. The durable question is not whether tokens become packets, but which network abstractions are stable enough to influence treatment without exposing model-specific internals.

INT CONDITIONAL - Importance, error tolerance, deadline, asymmetry, task context and service intent are more likely to become durable network objects than a direct dependency on one model's tokenization scheme. This reading would weaken if later specifications define model-specific token semantics as a first-class interoperability contract rather than using more general service abstractions.

INT WELL SUPPORTED - The first trajectory is AI-aware traffic. Earlier 6G material introduced token characteristics, error tolerance and importance as possible service inputs. By the SA2#176 cycle, architecture work was engaging with AI QoS, packet importance, asymmetric traffic and dynamic flow behavior. The important shift is not that “tokens” have become a universal packet type; it is that conventional traffic classes may not be sufficient to describe iterative, bursty and importance-sensitive AI workloads. [1],[9]

INT WELL SUPPORTED - A second trajectory is semantic representation. Service-level use cases describe sending task-relevant information instead of raw data; RAN1 material can use semantic language for compressed or learned physical-layer representations. Those are adjacent ideas, not interchangeable standards objects. A premium technical report should resist merging them simply because the word "semantic" appears in both. The implementation, performance metric and interoperability burden differ by layer. [1]

6.2 Semantic communication: useful direction, difficult boundary

INT WELL SUPPORTED - Semantic communication appears at several layers but should not be treated as one uniform concept. A service-level use case can describe transmitting task-relevant information rather than raw data, while a RAN1 discussion can use semantic ideas in learned compression or CSI representation. The scopes, evaluation metrics and implementation obligations differ. Treating them as synonyms would overstate convergence.

INT WELL SUPPORTED - The commercial attraction is clear: if a system can preserve task value while reducing bandwidth, latency or energy, semantic techniques may improve efficiency at the edge. The standards burden is harder. Meaning has to be measurable enough that sender and receiver share assumptions about reconstruction, quality, error tolerance and fallback. That challenge becomes more serious when the representation is learned and can change with model version or training data.

INT WELL SUPPORTED - The third trajectory is agent communication. A local agent may infer intent, delegate a task or coordinate with a server-side agent. Management agents may reason over observability data and translate operator intent into actions. Once an agent can change network state or invoke tools, discovery and messaging are no longer enough; authority, policy, auditability and failure containment become part of the protocol problem. [6],[7]

6.3 Agent-to-agent communication: from vocabulary to protocol study

SRC WELL SUPPORTED - The May baseline already observed agent-to-agent, MCP-like and intent-related proposals across RAN2, SA2, SA5 and CT. By September, TR 29.832 provides a dedicated Rel-20 draft study on protocol for AI in 6G. [1],[6] That institutional move is stronger evidence than the mere appearance of agent terminology because it creates an explicit place to resolve interoperability questions.

INT WELL SUPPORTED - The difficult issues are authority and context. An AI agent may discover another agent, delegate a task, call a tool, exchange context or act on a network resource. Each step can carry identity, privacy, authorization and audit implications. Mobile systems add mobility, session continuity and policy domains that generic Internet-agent protocols may not address in the same way. The 3GPP task is therefore likely to be selective profiling and system integration rather than reinventing every agent protocol.

INT WELL SUPPORTED - TR 29.832 is therefore significant because it creates a place where these questions can become protocol decisions rather than parallel vendor APIs. The study still has to determine boundaries: which functions 3GPP profiles itself, which external agent mechanisms it reuses, how identity and authorization map into mobile-system procedures, and whether mobility or network exposure creates 3GPP-specific requirements. [6]

INT WELL SUPPORTED - For operators, the commercial relevance is two-sided. AI-aware traffic could improve network treatment for valuable applications, but exposing model or task context can also create privacy, security and neutrality questions. Agentic control could reduce operational labor, but only if authority is bounded and the system remains auditable. The value proposition is therefore inseparable from governance.

6.4 Intent-driven operation and the problem of translation

INT WELL SUPPORTED - Intent is attractive because it lets an operator express a goal rather than a sequence of commands. The risk is translation: a high-level objective such as energy saving, reliability or service priority has to be converted into bounded actions across RAN, core and management domains. Different agents can interpret the same goal differently unless policy, observable state and allowed actions are explicit.

INT CONDITIONAL - Agentic operation can scale from recommendation toward bounded action only if authorization, policy validation, audit, conflict resolution, rollback and failure containment are explicit enough for carrier operations. The falsifier is an operating model in which an agent's action cannot be reconstructed, constrained or reliably reversed after it changes network state.

INT WELL SUPPORTED - The near-term watch is not whether every AI application receives a new QoS class. It is whether stable descriptors, lifecycle hooks and authorization semantics appear that application, core and RAN implementations can independently use. If they do, AI traffic and agent behavior begin to look like native system concerns rather than vendor extensions.

6.5 Protocol outlook: standardize the seams, not necessarily the intelligence

INT WELL SUPPORTED - A recurring theme across AI traffic, semantics and agents is that standards can define the seam without standardizing the internal intelligence. The network can specify what context is exchanged, how capability is advertised, how authorization is checked and what fallback applies while leaving model architecture and implementation proprietary. That pattern would preserve differentiation while reducing integration friction.

7 VERIFICATION: FROM TEST EVENT TO LIFECYCLE ASSURANCE

7.1 Why conventional RAN verification is not enough by itself

INT WELL SUPPORTED - Conventional telecom testing assumes that a defined implementation under defined configuration produces behavior that can be compared against a specification. Learned behavior introduces additional sources of variation: the model may differ by version, the data distribution may shift, the feature pipeline may change and local adaptation may modify state after deployment. A pass result therefore has to be interpreted in relation to the model and context that produced it.

INT WELL SUPPORTED - This does not make conformance impossible. It changes what evidence has to travel with the result. Applicability conditions, configuration, data collection, expected behavior and fallback can all become part of the assurance contract. The important design question is how much state must be exposed for a third party to reproduce or meaningfully compare behavior without requiring disclosure of proprietary model internals.

Figure 4. Verification moves from a test event to a lifecycle

Data	Model	Simulation	Digital twin	Lab / OTA	Field	Operations

Each step adds context that must remain traceable if a learned function is to be compared, replaced or safely updated.

INT WELL SUPPORTED - Traditional conformance assumes that the implementation under test is sufficiently defined that an input, configuration and expected behavior can be specified. Learned functions complicate that model because behavior depends on training distribution, model version, inference precision, feature extraction, update history and runtime. A passing result can therefore be fragile if the test never records the state that produced it.

INT WELL SUPPORTED - RAN5's growing involvement is a key signal because it forces those hidden assumptions into telecom artifacts. Applicability and PICS define when a feature should exist. Data-collection procedures define what inputs are available. Test cases have to establish repeatability without making the standard accidentally depend on one proprietary model. The challenge is to test behavior while preserving implementation freedom. [5],[28]

7.2 The RAN5 precedent and its limits

SRC WELL SUPPORTED - RAN5 already works with AI/ML applicability, PICS, test cases and data-collection changes for current-generation features. [5],[28] The direct evidence should be scoped carefully: this is not a completed 6G AI conformance architecture. Its importance is that 3GPP has an institutional mechanism for turning AI functionality into testable telecom artifacts.

INT WELL SUPPORTED - The precedent matters for procurement because it creates a language in which a buyer can ask what implementation conditions were tested, what was observed and what version or configuration produced the result. As AI moves into more features, the reusable part may be less a single test case than the meta-structure for expressing applicability, state and expected behavior.

CTX WELL SUPPORTED - The Keysight/Samsung/NVIDIA validation workflow is relevant because it connects scenario and data generation, model training and benchmarking. It is still an ecosystem demonstration rather than an industry-wide conformance regime, but it illustrates the direction verification must take: the test environment becomes part of the AI lifecycle, not merely the final RF gate. [22]

7.3 Generalization, drift and repeatability

INT WELL SUPPORTED - A model can pass a carefully constructed test and still fail to generalize across geography, band, device mix, mobility, traffic or interference. The relevant telecom question is not whether a model is universally correct, but whether its operating envelope is known and whether the system can detect when it leaves that envelope. That makes monitoring and fallback part of verification rather than purely operational afterthoughts.

INT WELL SUPPORTED - Model drift adds time to the assurance problem. A function that was valid at acceptance can change after a software update, retraining event or environmental shift. Continuous monitoring therefore needs a reference: performance expectations, model identity, data conditions and update history. Without that traceability, operators can observe a KPI change without being able to attribute it to radio conditions, model behavior or deployment configuration.

INT WELL SUPPORTED - The hardest question is generalization. A model can outperform a baseline in one channel, mobility pattern or traffic mix and degrade elsewhere. Verification must therefore distinguish feature correctness from model quality and deployment suitability. Conformance can establish that standardized behavior is implemented; it may not prove that the model is optimal or safe in every operator environment.

INT WELL SUPPORTED - This creates a new procurement opportunity for test and measurement vendors, OTICs, operator labs and digital-twin providers. The premium value is not simply more test cases. It is traceability: which dataset, model, runtime and environment produced the result, and whether the same evidence can be regenerated after an update.

7.4 Digital twin, lab, OTA and field evidence serve different purposes

INT WELL SUPPORTED - Simulation and digital twins are strong for controlled repeatability and rare scenarios; lab and OTA systems add hardware and radio realism; field trials add topology, mobility and operational complexity; commercial operation reveals lifecycle cost and sustained behavior. No single layer dominates the others. A credible assurance strategy uses them as a chain in which each stage closes a different uncertainty.

INT WELL SUPPORTED - The most valuable result is therefore not the largest isolated KPI gain. It is a result that survives movement down the evidence chain with enough metadata to explain why. That is the difference between a research benchmark and a product claim an operator can use in acceptance, regression or procurement.

INT WELL SUPPORTED - A practical verification doctrine for AI-RAN is therefore layered: validate the data, validate the model, validate the function, validate the RF/system interaction, validate fallback, and continue monitoring after deployment. The lifecycle does not end at acceptance because retraining or model replacement can change behavior without changing the nominal feature name.

7.5 Verification as interoperability and switching-cost control

INT WELL SUPPORTED - The verification problem is also a market-structure problem. If only the incumbent vendor can reproduce a model's behavior because critical data, state or evaluation assumptions are hidden, switching cost rises even when network interfaces are formally open. Conversely, shared assurance semantics can let competing implementations prove behavior without exposing every proprietary detail. The model may remain closed; the evidence required to trust, compare and replace it need not be.

8 INDUSTRY MOMENTUM AND COMPETING COMPUTE PATHS

8.1 Ecosystem scale is no longer the main question

CTX ESTABLISHED - As of 26 February 2026, the AI-RAN Alliance reported 132 members, 33 MWC 2026 demonstrations and four Industry Blueprints. [13] Those numbers establish breadth of interest at that snapshot; they do not establish September membership, operator deployment consensus or the commercial superiority of any one architecture.

INT WELL SUPPORTED - The more useful 2026 question is how evidence is migrating outward. The field now includes alliance blueprints, vendor platforms, live-network AI-for-RAN trials, dynamic orchestration demonstrations, external workload experiments, physical-AI proofs and dedicated validation workflows. That progression shows technical breadth but also increasing heterogeneity: different demonstrations close different risks and should not be rolled into one maturity score.

SRC ESTABLISHED - The AI-RAN Alliance reported 132 members and 33 MWC 2026 demonstrations, a useful indicator that the ecosystem has moved beyond a small research circle. The more important change is qualitative: activity now spans AI-for-RAN optimization, shared AI/RAN infrastructure, AI workloads on RAN-adjacent compute, data, testing, digital twins, agentic networking and commercialization. [13]

CTX WELL SUPPORTED - Nokia and NVIDIA have the most visible accelerator-centric commercial story. Nokia's September update named multiple operators working on PoCs or live trials and reported spectral-efficiency improvements on its platform. Those results are vendor reported and should be treated as such, but the breadth of named operator activity is meaningful because it moves the proposition from a bilateral partnership toward an ecosystem test. [14]

8.2 Nokia/NVIDIA: richer common infrastructure and platform economics

CTX WELL SUPPORTED - Nokia's September announcement expands the accelerator-centric thesis through operator PoCs and live trials on NVIDIA platforms and reports more than 20% spectral-efficiency improvement in the cited vendor context. [14] The strategic proposition is broader than a single radio algorithm: a common accelerated platform could support RAN processing, AI workloads and potentially external edge applications.

INT WELL SUPPORTED - That architecture has a clear upside if high utilization can amortize expensive compute and if a mature AI software ecosystem accelerates deployment. It also concentrates several proof burdens: real-time isolation, peak power, cooling, orchestration, transport, failure behavior and the cost of keeping accelerator capacity economically busy when radio load and external workload demand vary independently.

CTX WELL SUPPORTED - Ericsson provides an important counterexample. Its 2026 AI-in-RAN strategy emphasizes software functions running on existing RAN compute and cites more than 15 deployments or trials. T-Mobile and Ericsson also reported live-network gains for AI-native scheduling/link adaptation. That evidence does not disprove accelerated common infrastructure; it demonstrates that meaningful AI-for-RAN value can emerge without first adopting a new universal compute substrate. [15],[16]

8.3 Ericsson and Samsung: AI-native behavior without one mandatory accelerator

CTX WELL SUPPORTED - Ericsson's June 2026 position emphasizes AI in RAN running on Ericsson baseband/radio/RAN compute without requiring a new universal GPU platform, while Samsung explicitly states that AI does not equal GPU and frames CPU/GPU selection around service requirements and economics. [15],[25] These are vendor positions, not standards outcomes, but together they demonstrate that AI-native behavior and accelerator standardization are separate questions.

INT WELL SUPPORTED - The existence of credible alternative compute paths is strategically healthy for the analysis because it creates a natural experiment. If similar AI-for-RAN gains can be achieved with materially different hardware and software stacks, the market can compare performance, power, cost and portability rather than assuming one implementation model from the outset.

CTX WELL SUPPORTED - SoftBank makes the landscape more interesting still. Its AITRAS program tests common infrastructure, orchestration and external AI workloads, while separate Ericsson work explores interworking and physical-AI use cases. The commercial question becomes whether the same operator can use different implementation models at different sites or for different purposes rather than choosing one architecture for the entire network. [18]-[20],[30]

INT WELL SUPPORTED - This diversity is commercially healthy. Macro sites, enterprise campuses, regional edge locations and dense urban clusters do not have identical power, transport or workload economics. A high-value industrial site may justify richer accelerated compute long before the median macro cell. Conversely, software AI on existing compute may scale broadly when the objective is capacity or quality rather than third-party workload hosting.

8.4 SoftBank: operator portfolio rather than single-path architecture

CTX WELL SUPPORTED - SoftBank's public work spans AITRAS commercialization, external AI workloads with Nokia, orchestration interworking with Ericsson, Physical-AI PoCs and live-network AI-in-RAN trials. [17]-[20b] The breadth is useful precisely because it separates functions that are often bundled rhetorically: network optimization, shared infrastructure, orchestration and external application support.

INT WELL SUPPORTED - This suggests a portfolio interpretation of AI-RAN. Operators may not need one architecture for every site or business. Dense urban or edge locations can justify richer compute; broad macro coverage can use software AI on existing platforms; private or industrial sites can support locally valuable inference; external workloads can remain selective until demand becomes repeatable. Heterogeneity can therefore be an economic design choice, not an implementation failure.

INT WELL SUPPORTED - The industry signal is therefore not “one camp has won.” It is that AI is forcing the RAN stack to negotiate with compute economics more explicitly than before. That makes portability, orchestration and proof critical because operators may want heterogeneous architectures by geography, workload and lifecycle stage.

8.5 Competitive boundary: platform convergence is not implied by standards convergence

INT WELL SUPPORTED - The industry evidence reinforces the standards reading. 3GPP is increasingly specific about behavior, data, lifecycle, service awareness and proof, while vendors remain free to differentiate compute placement and runtime architecture. The likely competitive battleground is therefore a layered one: open or standardized contracts above, differentiated silicon and software underneath, and operator procurement deciding how much coupling is acceptable.

9 AI-FOR-RAN, AI-AND-RAN AND AI-ON-RAN: THREE COMMERCIAL THESES

9.1 AI-for-RAN: the shortest path from technical gain to operator value

CTX WELL SUPPORTED - T-Mobile and Ericsson reported close to 10% spectral-efficiency improvement and up to 15% downlink-throughput improvement in a live commercial-network trial; SoftBank and Ericsson later reported approximately 10% average improvement across evaluated locations, with higher site-specific peaks. [16],[17] These results are attributed to the reporting companies and are not normalized as a cross-vendor benchmark.

INT WELL SUPPORTED - AI-for-RAN has the clearest business path because the benefit can remain inside the network budget. A persistent capacity gain can defer spectrum or site expansion; better energy behavior can reduce operating cost; improved link adaptation can improve experience without a new external buyer. The remaining proof is persistence across topology, traffic, mobility and software versions, plus the compute and lifecycle cost required to sustain the gain.

INT WELL SUPPORTED - AI-for-RAN has the most direct economic chain because the buyer is already the operator and the KPI is already part of network planning. If an AI scheduler improves spectral efficiency or throughput across a representative network mix with acceptable power and stability, the benefit can be translated into capacity, user experience or cost deferral. This is why 2026 live-network evidence from Ericsson/T-Mobile and SoftBank/Ericsson matters even when the exact results are not directly comparable. [16],[17]

INT WELL SUPPORTED - AI-and-RAN asks a harder question: can the operator run radio processing and other AI workloads on a common accelerated platform without compromising deterministic RAN behavior? The technical demonstration is only the beginning. Carrier isolation, scheduling, telemetry, failure domains, cooling, peak power, transport placement and real utilization determine whether shared infrastructure creates value rather than simply moving cost. [18]-[20]

9.2 AI-and-RAN: shared infrastructure is an economic thesis

INT WELL SUPPORTED - AI-and-RAN proposes that radio processing and AI workloads share richer compute infrastructure so that capacity that would otherwise be idle can create additional value. The technical challenge is to preserve carrier timing, reliability and isolation while dynamically allocating resources. The commercial challenge is utilization: an expensive common platform only improves economics if the workload mix is dense and predictable enough to justify power, cooling and orchestration overhead.

CTX WELL SUPPORTED - SoftBank's AITRAS work and orchestration demonstrations are useful evidence of the architecture direction, and Nokia/NVIDIA operator trials increase the number of sites at which the thesis can be tested. [14],[18],[20a],[30] The missing evidence is still a broad multi-site economic record, not another demonstration of functional feasibility.

INT WELL SUPPORTED - AI-on-RAN is commercially the most ambitious thesis because it assumes the operator can expose distributed AI capability to applications or enterprises. Physical-AI pilots, video analytics and external workload brokering show technical plausibility. They do not yet establish a large category of customers who repeatedly buy edge inference from mobile operators instead of using on-premise or hyperscale alternatives. [19]-[21]

9.3 AI-on-RAN: the network as a distributed AI service platform

CTX WELL SUPPORTED - External workload and Physical-AI programs from SoftBank/Nokia, SoftBank/Ericsson and NVIDIA/T-Mobile show a technically plausible path in which distributed network infrastructure hosts inference close to devices and radio context. [19]-[21] These programs are best read as ecosystem and PoC evidence, not as proof of a repeatable operator revenue model.

INT WELL SUPPORTED - The commercial hurdle is higher because the buyer is no longer only the network organization. Enterprises and application developers already have hyperscale cloud, regional edge and on-premises alternatives. Operator-hosted AI therefore needs a reason to exist: radio-aware locality, mobility, deterministic service treatment, sovereignty, integrated sensing or an economic advantage that those alternatives cannot provide as efficiently.

INT WELL SUPPORTED - The three theses can coexist. An operator may deploy AI-for-RAN broadly, use accelerated common infrastructure selectively where utilization is high, and offer AI-on-RAN only in locations with anchor customers. That heterogeneous outcome is more consistent with current evidence than a uniform "GPU at every site" or "AI service at every cell" narrative.

INT WELL SUPPORTED - This maturity separation also clarifies what to watch. For AI-for-RAN, look for fleet expansion, energy data and persistent KPI gains. For AI-and-RAN, look for real utilization, SLA isolation and TCO. For AI-on-RAN, look for named paying customers, billing models and repeat workload density. The proof object changes as the thesis changes.

9.4 Why the three theses should not be scored as one market

INT WELL SUPPORTED - AI-for-RAN can succeed even if shared GPU infrastructure never becomes economical at ordinary sites. AI-and-RAN can create infrastructure value even if third-party AI demand remains limited. AI-on-RAN can find attractive industrial niches without becoming a mass-market network service. Treating all three as one AI-RAN maturity curve hides the fact that they have different buyers, evidence thresholds and failure modes.

Figure 5. Commercial proof is a ladder, not a label

Concept	PoC	Field / live	Deployment	Repeat revenue

A live-network result closes a different risk from a deployed product, and neither proves repeatable external revenue. The report keeps those stages separate.

9.5 The 5G lesson revisited: repeatability closes more than novelty

INT WELL SUPPORTED - Commercial readiness is less about whether a capability is impressive than whether its benefit survives repetition. For AI-for-RAN that means persistent KPI and power results. For shared infrastructure it means stable utilization and carrier isolation. For external AI it means repeat customers and margins. The evidence ladder is therefore a discipline against extrapolating from one successful PoC to an assumed fleet-wide business.

10 OPERATOR EVIDENCE, VERTICALS, ISAC AND PHYSICAL AI

10.1 Operators as the reality check on vendor architecture

INT WELL SUPPORTED - Operators carry the lifecycle cost that architecture diagrams do not show. They pay for site power, cooling, transport, integration, software maintenance, security, field operations and capacity planning. Their trial behavior therefore provides a different kind of evidence from vendor product announcements. An operator that evaluates multiple approaches is not necessarily indecisive; it may be separating site classes, use cases and economic thresholds before committing to scale.

INT WELL SUPPORTED - T-Mobile is particularly informative because it participates in both Nokia/NVIDIA and Ericsson initiatives. SoftBank is similarly informative because it spans AITRAS, Nokia external-workload work, Ericsson orchestration, Physical AI and live-network AI-in-RAN. The common lesson is plurality: operator experimentation in 2026 is broader than a single-vendor alliance map would suggest.

INT WELL SUPPORTED - T-Mobile has become one of the most visible AI-RAN laboratories because it works across both Nokia/NVIDIA and Ericsson initiatives. That is strategically useful: the same operator can compare accelerator-centric infrastructure with AI features running on existing RAN platforms rather than evaluating each architecture in isolation. The evidence suggests experimentation with multiple paths, not a completed selection of one universal stack. [14],[16],[21]

INT WELL SUPPORTED - SoftBank offers another multi-path case. AITRAS explores common infrastructure, orchestration and external workloads, while work with Ericsson tests AI-in-RAN and physical-AI interworking. The commercial lesson is that operator architecture can be portfolio-based: one model for high-value edge sites, another for broad macro coverage, and different vendors around a shared lifecycle or orchestration layer. [17]-[20],[30]

CTX WELL SUPPORTED - The September U.S. operator discussion is also instructive, even though trade-press interviews are weaker evidence than standards records or deployment data. Public comments from AT&T;, Verizon and T-Mobile show different views on the role of GPUs in radio processing versus edge applications and on the economics required to justify common accelerated infrastructure. This divergence is precisely why "AI-RAN adoption" should not be reduced to a single hardware indicator. [24]

10.2 Manufacturing and robotics: where locality and control can justify edge intelligence

INT WELL SUPPORTED - Manufacturing and robotics are credible verticals when network locality materially changes the outcome. Machine vision, robot coordination, predictive control and safety monitoring can benefit from low latency, local data handling and predictable service. The relevant question is not whether a factory can run AI - it already can - but whether integrating communication, mobility and distributed inference creates enough operational advantage to justify operator or private-network infrastructure.

CTX WELL SUPPORTED - Physical-AI PoCs from SoftBank/Ericsson and NVIDIA/T-Mobile provide evidence that network-enabled robotics, video analytics and industrial applications can be integrated with AI-RAN-ready infrastructure. [20b],[21] They do not yet establish repeatable deployment economics, safety assurance or a standard commercial model.

INT WELL SUPPORTED - Manufacturing and robotics. The strongest case appears where low-latency control, local inference, deterministic connectivity, mobility and data sovereignty all matter at once. Physical-AI and industrial edge pilots can benefit from radio awareness and local compute, but safety assurance and systems integration are at least as important as inference latency. A plant that already owns private edge infrastructure may not need an operator AI platform unless the network provides a measurable integration advantage. [20],[21]

INT WELL SUPPORTED - Video analytics and smart-city environments. High uplink volume and location-specific inference create an intuitive edge case. Network-assisted placement can reduce backhaul and response time, and agentic search can turn continuous streams into task-driven workflows. Yet privacy, GPU utilization and the buyer model remain constraints. A city or venue may prefer dedicated compute unless the operator can bundle connectivity, locality and lifecycle management at a compelling price.

10.3 Video analytics, smart city and public-space sensing

INT WELL SUPPORTED - High-uplink video and event-driven inference create an intuitive role for distributed compute because raw streams can be bandwidth-intensive and local response can matter. The network can potentially add mobility, priority, location and radio-state information that a generic edge platform does not have. The counterweight is privacy and governance: richer sensing and context can increase the sensitivity of the data as quickly as it increases technical value.

INT WELL SUPPORTED - The business case depends on who buys the integrated outcome. A municipality, venue, transport operator or enterprise may value an end-to-end service that combines connectivity and analytics, but the operator has to compete with specialist video platforms and private infrastructure. Network advantage therefore has to be demonstrated in reliability, coverage, integration or cost rather than assumed from physical proximity alone.

INT WELL SUPPORTED - Automotive and V2X. Mobility, positioning, cooperative perception and sensing align naturally with AI-assisted radio and ISAC. The strategic value is not only lower latency but shared situational context across vehicles and infrastructure. The commercial hurdle is reliability and liability: a sensing-assisted service is useful only if coverage continuity, calibration and failure behavior are explicit enough for the application to trust. [12],[29]

10.4 Automotive, drones and low-altitude systems: sensing and mobility converge

INT WELL SUPPORTED - Automotive and low-altitude systems combine mobility, positioning, perception, uplink media and safety-sensitive control. They are natural places to connect AI-RAN with integrated sensing and communication because the network can become both a communications fabric and an environmental observation layer. The value is strongest when network-assisted perception or positioning improves decisions that a device cannot make as reliably alone.

SRC WELL SUPPORTED - ITU-R M.2160 includes Artificial Intelligence and Communication and Integrated Sensing and Communication among IMT-2030 usage scenarios, while 3GPP RAN work continues to progress NR sensing / ISAC studies. [12],[29] The standards trajectory supports the capability direction; it does not by itself prove a vertical business case or resolve privacy and liability.

INT WELL SUPPORTED - Ports, logistics and utilities. Wide-area industrial mobility, robots, inspection video and remote assets create a mixed private/public network environment. Edge inference can reduce bandwidth and improve autonomy, while AI-for-RAN can optimize coverage and quality. The decision is often operational rather than technological: who integrates the site, who owns the data, and who carries responsibility when an automated system acts on incomplete network information?

INT WELL SUPPORTED - Enterprise private networks and XR. Private networks benefit from predictable locality and sovereignty; XR adds throughput, positioning and rendering sensitivity. Both can use network-aware AI without requiring a new public-network revenue model. They are therefore useful proving grounds for AI-native behavior, even if the business remains enterprise infrastructure rather than “AI-on-RAN as a service.”

10.5 Enterprise, ports, utilities and remote systems: selective deployment economics

INT WELL SUPPORTED - Private networks, ports, utilities and remote industrial systems can justify richer local compute when data sovereignty, resilience, uplink demand or operational autonomy matters. These environments may be better early targets for integrated AI/network platforms than a uniform macro-network rollout because the value per site is higher and the workload can be more predictable.

INT WELL SUPPORTED - The same logic cautions against extrapolating niche success to fleet-wide deployment. A high-value factory can support local acceleration, dedicated integration and specialist operations that would be uneconomic across thousands of ordinary macro sites. The likely near-term market therefore fragments by site class and vertical rather than converging on a single infrastructure template.

INT WELL SUPPORTED - Drones, low-altitude systems and NTN. Mobility, sensing, intermittent links and constrained energy make selective data reduction and onboard inference especially relevant. Here the network role may be coordination and awareness rather than raw compute hosting. AI-assisted scheduling, compression and sensing could matter even where inference remains on the device or satellite. The common thread across these verticals is not “edge AI”; it is that network state materially changes application outcome.

10.6 ISAC as a multiplier, not a guaranteed product

INT CONDITIONAL - If sensing data becomes available to applications under standardized controls, the network can move from better connectivity toward network-assisted perception. That can increase value in robotics, transport, inspection and spatial computing. The condition is trustworthy calibration, privacy controls, integration and liability allocation. Without those, sensing may remain a technical capability whose commercial use is narrower than its architectural importance.

11 COMMERCIAL READINESS AND OPERATOR ECONOMICS

11.1 Performance is necessary; economics is decisive

INT WELL SUPPORTED - AI-RAN enters a mature and cost-sensitive RAN market. New value has to come from better spectrum use, lower operating cost, higher infrastructure utilization or new services. Each route has a different accounting logic. A radio KPI improvement can have value through deferred capacity; an energy feature through OPEX; shared infrastructure through utilization; external AI through customer revenue. A generic statement that AI improves the network is therefore commercially incomplete.

INT WELL SUPPORTED - The most useful economic question is marginal: what additional compute, memory, power, cooling, transport, license, lifecycle and operational burden is required to obtain the measured gain? A feature that produces 10% performance improvement with modest incremental cost has a different deployment profile from one that requires a new accelerated platform at every site. Public technical results rarely expose the full denominator, which is why the report avoids converting KPI claims into ROI estimates.

INT WELL SUPPORTED - The traditional RAN market is not a high-growth blank slate. New AI-RAN value therefore has to come from one or more of four sources: better use of spectrum, lower operating cost, higher infrastructure utilization or new services. Each source has a different evidence threshold. The first two can be justified inside the network budget; the latter two require a stronger platform and demand thesis.

INT WELL SUPPORTED - AI-native scheduling and link adaptation currently have the clearest readiness profile. The evidence includes live commercial-network results and multiple vendor/operator trials. The missing commercial proof is persistence: do gains hold across bands, traffic mixes, mobility and software versions, and what is the power delta required to sustain them? [15]-[17]

INT WELL SUPPORTED - Other AI-for-RAN functions can also create value through optimization, energy saving and automation, but the market needs comparable benchmarks and failure/fallback evidence. Vendor claims are useful signals; they are not yet a common yardstick. A premium procurement decision will ask for the baseline, test environment, confidence range and operational cost of keeping the model healthy.

11.2 AI-for-RAN readiness: strongest evidence, still context-dependent

INT WELL SUPPORTED - Live-network scheduler and link-adaptation evidence makes AI-for-RAN the most commercially mature of the three AI-RAN theses. The missing proof is breadth rather than existence. Operators need to know whether gains persist across band, topology, mobility, traffic mix, device population and software evolution, and whether lifecycle cost remains acceptable when models are monitored and updated over years rather than weeks.

INT WELL SUPPORTED - A feature can also be valuable without being universal. If AI scheduling works best in dense cells, high-traffic bands or certain hardware generations, targeted deployment may still produce attractive economics. Commercial maturity should therefore not be equated with identical rollout across every site. Selectivity can be a rational optimization rather than a sign that the technology failed to scale.

INT WELL SUPPORTED - GPU or common accelerated RAN has a different burden. The investment case is not “GPU is faster.” It is whether shared infrastructure can deliver RAN performance plus enough AI utilization to offset capex, power, cooling and site complexity. That proof must be site-specific because dense urban edge locations, industrial campuses and conventional macro sites have different workload density and physical constraints. [14],[18],[19]

11.3 Common accelerated infrastructure: utilization is the key variable

INT WELL SUPPORTED - The common-platform thesis is sensitive to utilization because accelerator economics worsen quickly when expensive capacity sits idle. RAN load follows daily and geographic patterns; external AI load follows different demand curves. Orchestration can improve sharing, but it also has to preserve radio deadlines and isolation under contention. The business case therefore depends on the joint distribution of workloads, not simply the peak performance of the accelerator.

INT WELL SUPPORTED - Site diversity matters as well. Power, cooling, transport and physical space vary substantially across central offices, regional edge sites, enterprise facilities and macro locations. A platform that works well in a data-center-like edge site can be uneconomic at a constrained cell site. This is the basis for the report's selective-acceleration outlook.

INT WELL SUPPORTED - Dynamic AI/RAN orchestration adds another layer. A scheduler that can shift compute between telecom and external AI is technically interesting, but its commercial value depends on SLA semantics, telemetry, portability and customer demand arriving at the right time and place. Spare capacity has value only when there is a buyer whose workload can safely use it. [18]-[20],[30]

11.4 External AI revenue: capability is not demand

INT WELL SUPPORTED - The technical ability to execute external AI workloads on telecom infrastructure is a necessary but weak commercial signal. A stronger signal is a named application that repeatedly chooses operator-hosted inference because locality, radio awareness, sovereignty or integrated service quality creates measurable value. Stronger still are contracted workload volume, SLA terms, billing and margin. Public 2026 evidence is still concentrated in PoCs, ecosystem programs and commercialization narratives rather than a broad revenue base.

INT WELL SUPPORTED - This does not make the thesis unimportant. New platforms often mature through narrow verticals before becoming general services. The correct reading is conditional: external AI becomes strategically larger if repeat demand appears in places where the operator has a structural advantage. If customers continue to prefer hyperscale, private or on-premises alternatives, AI-on-RAN can remain technically capable but commercially specialized.

INT WELL SUPPORTED - Physical AI and external workloads therefore remain earlier on the business curve. The next decisive evidence is named customers, repeat deployments, billing, workload density and margin. A pilot proves that the operator can host the workload. Repeatable revenue proves that the operator should.

INT WELL SUPPORTED - The most credible near-term pattern is selective deployment. High-demand urban sites, enterprise locations and industrial edge clusters may justify richer compute earlier than the median macro site. A uniform all-sites accelerator thesis requires substantially stronger power/TCO evidence than the public record currently provides.

11.5 Procurement and policy: control surfaces become economic variables

INT WELL SUPPORTED - Procurement increasingly has to ask not only whether an interface is standardized but whether models, data, telemetry, updates and rollback can be governed independently of the incumbent vendor. A low headline compute price can be offset by switching cost if lifecycle state is not portable. Conversely, a more expensive platform can be attractive if it reduces integration, assurance and operational burden. The relevant economic unit is therefore the lifecycle, not only the accelerator.

12 FINDINGS, CONDITIONAL OUTLOOK AND WATCHLIST

12.1 Finding: AI-native 6G is more radical in information than in structural split

INT WELL SUPPORTED - Madrid's HLS baseline weakens the idea that AI-native 6G must begin with a clean-sheet higher-layer architecture. The deeper change is that information previously treated as implementation detail - model state, data conditions, intent, telemetry, sensing and assurance - can become relevant to interoperable behavior. That is a substantial architectural shift even when familiar CU-DU and CP-UP boundaries remain.

12.2 Finding: the institutional footprint of AI is expanding faster than any one algorithm

INT WELL SUPPORTED - RAN, RAN5, SA2, SA5, SA6 and CT3 now each own a different part of the learned-system problem. That institutional distribution is itself a maturity signal because it means AI behavior is being translated into ordinary telecom responsibilities: radio procedures, service architecture, management, security, application enablement, protocol and test. The strategic value lies in the interfaces between those responsibilities.

12.3 Finding: verification is becoming part of product architecture

INT WELL SUPPORTED - If learned behavior cannot be reproduced, bounded and updated safely, it cannot be treated like a durable carrier feature. Verification therefore moves upstream into design: model and data state have to be considered while interfaces and lifecycle are being defined, not only after the implementation is complete. This makes assurance a potential interoperability surface and a procurement tool.

12.4 Finding: platform lock-in may migrate upward

INT WELL SUPPORTED - Traditional standards can preserve radio-level openness while AI creates new switching friction above the interface. Feature pipelines, runtime libraries, quantization, telemetry, data schemas, orchestration and agent frameworks can become difficult to replace even when packets and procedures are standardized. The important policy and procurement question is therefore which assumptions are becoming invisible dependencies.

12.5 Finding: heterogeneous deployment is a plausible equilibrium

INT WELL SUPPORTED - The 2026 evidence does not require a future in which every site converges on the same accelerated compute architecture. Broad coverage can use embedded or software AI, high-value sites can justify richer acceleration, and external AI can remain concentrated where application demand exists. This topology can persist if site economics and workload density remain heterogeneous.

INT WELL SUPPORTED - Finding 1 - AI-native 6G looks evolutionary in structure and more radical in information. Madrid's HLS baseline reduces the case for a clean-sheet structural narrative. The more consequential change is that models, data, sensing, intent and lifecycle are entering system design as ordinary objects. [2]

INT WELL SUPPORTED - Finding 2 - the institutional footprint of AI is expanding faster than any single AI algorithm. RAN, RAN5, SA2, SA5, SA6 and CT3 now each own part of the problem. That makes cross-layer contracts more strategically important than a headline model architecture. [4]-[10]

INT WELL SUPPORTED - Finding 3 - verification is becoming part of product architecture. AI-RAN cannot be scaled safely if model state, data assumptions and fallback behavior are invisible to acceptance testing. RAN5 and digital-twin work are therefore leading indicators of commercial maturity, not downstream housekeeping. [5],[22],[23],[28]

INT WELL SUPPORTED - Finding 4 - AI-for-RAN is likely to mature before the broader shared-compute thesis. Its value can be captured through capacity, quality, energy or automation without waiting for external AI revenue. The evidence clock is furthest right for scheduling/link adaptation and related optimization. [15]-[17]

INT WELL SUPPORTED - Finding 5 - AI-and-RAN is becoming an economics experiment. The technology question is largely plausible; the harder proof is whether real sites have enough utilization, power headroom and workload demand to justify the common platform. [18]-[20]

INT WELL SUPPORTED - Finding 6 - physical AI is strategically interesting because it connects network locality to a new demand curve. But the category becomes commercially meaningful only when operators can show named, repeat customers rather than ecosystem demonstrations. [20],[21]

INT WELL SUPPORTED - Finding 7 - agentic networking raises governance stakes. Once AI can translate intent into network action, discovery and messaging are insufficient. Authorization, policy, audit and safe rollback become protocol and operations requirements. [6],[7]

INT WELL SUPPORTED - Finding 8 - platform lock-in may migrate upward. Even with standardized radio architecture, dependence can accumulate in model lifecycle, data schemas, runtime libraries, observability and orchestration. [1]

INT WELL SUPPORTED - Finding 9 - 2027 will be a conversion year. The question will be which Rel-20 studies and agreements are concrete enough to become Rel-21 normative scope, and which remain useful research vocabulary without interoperable contracts. [11]

INT WELL SUPPORTED - Finding 10 - the decisive evidence is increasingly operational. Power/TCO, fleet-scale repeatability, multi-vendor portability, conformance and customer revenue will matter more than another demo count. That is where standards, market and evidence clocks begin to meet.

12.6 Layer 1 outlook: what follows if current consensus continues to harden

INT CONDITIONAL - **Observation:** 5G HLS remains the 6G RAN HLS baseline while AI/data/agent work broadens. **Condition:** selected Rel-20 studies convert into stable Rel-21 scope and assurance machinery continues to mature. **Consequence:** 6G becomes structurally evolutionary but operationally more AI-native, with novelty concentrated in lifecycle, observability and control. **Falsifier:** later Rel-21 decisions materially reopen protocol distribution or replace the HLS baseline.

INT CONDITIONAL - **Observation:** current NR AI/ML work already produces applicability, PICS, test and data-collection artifacts. **Condition:** reusable assumptions and traceable model/data state emerge across more features. **Consequence:** assurance becomes a procurement surface for learned implementations. **Falsifier:** evidence remains too feature-specific or runtime-specific to support meaningful comparison.

INT CONDITIONAL - **Observation:** TR 29.832 and adjacent architecture work create a formal place for AI protocol questions. **Condition:** identity, authorization, discovery and context exchange converge on reusable semantics. **Consequence:** mobile-system agent interoperability can harden without standardizing the agent's internal model. **Falsifier:** the study remains descriptive or incompatible external assumptions dominate implementation.

INT WELL SUPPORTED - Layer 1 - architecture. If RAN3 continues to preserve the 5G protocol distribution while optimizing around it, Rel-21 is more likely to concentrate architectural novelty in new functions, data and AI semantics than in a wholesale higher-layer split redesign. This reading weakens if a later plenary reopens protocol distribution or adopts a materially different HLS. [2]

INT WELL SUPPORTED - Layer 1 - AI/ML framework. If RAN3 turns 6G use cases and data-framework discussion into stable information exchanges, model/data exchange and lifecycle hooks could become native 6G functions even without standardizing the model itself. This weakens if the AI scope remains implementation-specific and no common interfaces emerge.

INT WELL SUPPORTED - Layer 1 - protocol and agents. If TR 29.832 converges on reusable communication, discovery and authorization primitives, 6G could standardize profiles for agent/service interaction rather than rely on ad hoc APIs. This weakens if the study closes without actionable scope or defers completely to mechanisms that do not map cleanly into mobile-system behavior. [6]

INT WELL SUPPORTED - Layer 1 - conformance. If RAN5 expands beyond narrow AI cases into reusable model-aware test patterns, procurement may increasingly require explicit model/data state and fallback behavior. This weakens if testing remains feature-specific and operators continue to rely mainly on proprietary acceptance methods. [5],[28]

12.7 Layer 2 outlook: downstream possibilities that require an additional condition

INT CONDITIONAL - **Portable AI lifecycle.** Layer-1 prerequisite: intelligence interfaces stabilize. Additional condition: model-state, version, data, feedback and fallback semantics become implementation-independent enough for substitution. Consequence: multi-vendor portability extends beyond protocol compatibility. Falsifier: proprietary model/runtime coupling remains a blocking dependency.

INT CONDITIONAL - **Selective accelerated infrastructure.** Layer-1 prerequisite: AI-for-RAN gains scale. Additional condition: common-platform economics are compelling at only some site classes. Consequence: embedded/software AI dominates broad coverage while GPU/NPU infrastructure concentrates at high-value sites. Falsifier: ordinary macro-site economics support broad uniform acceleration, or accelerated sites still fail to achieve sufficient utilization.

INT CONDITIONAL - **AI-on-RAN service business.** Layer-1 prerequisite: distributed AI capability is technically reliable. Additional condition: named paying customers, repeat workloads, SLA terms, billing and margins appear at meaningful scale. Consequence: distributed AI becomes a distinct operator service layer. Falsifier: evidence remains project-specific and customers prefer hyperscale, regional cloud or private alternatives.

INT CONDITIONAL - **Bounded agentic operations.** Layer-1 prerequisite: agents can be discovered and communicate. Additional condition: authorization, policy, audit, rollback and failure containment become explicit enough for carrier control. Consequence: agentic operation can move from recommendation toward bounded action. Falsifier: operators cannot reliably reconstruct or reverse agent actions after state changes.

INT WELL SUPPORTED - Layer 2 - heterogeneous compute. If the 5G HLS baseline persists and cloud/edge transport economics improve, more DU/CU/AI functions could migrate toward regional or edge compute without requiring a clean-sheet split. The boundary is topology: fronthaul, latency and power economics vary enough that dedicated hardware may remain cheaper or more reliable at many sites.

INT WELL SUPPORTED - Layer 2 - deployment mix. If AI-for-RAN gains scale and common-accelerator economics remain weak at ordinary sites, embedded/software AI could dominate broad-area deployment while GPU-rich AI-RAN concentrates at high-value locations. The forecast reverses if accelerated platforms become sufficiently cheap and power-efficient for routine macro rollout.

INT WELL SUPPORTED - Layer 2 - edge-AI business. If AI/RAN orchestration matures and spare compute attracts external workload demand, selected operators could build an edge-AI utilization business around RAN locations. The boundary is demand density: workloads may still prefer hyperscale, regional cloud or customer-owned private edge.

INT WELL SUPPORTED - Layer 2 - physical AI. If industrial pilots prove low latency and reliability and safety/integration frameworks mature, manufacturing, robotics and video analytics could become repeatable AI-on-RAN verticals. The forecast weakens if customers deploy equivalent capability on private on-premise infrastructure without operator involvement.

INT WELL SUPPORTED - Layer 2 - agentic operations. If agent protocol and OAM semantics standardize and authorization remains operator-controlled, agentic operations could reduce manual work without surrendering policy control. The forecast weakens if agents cannot be bounded, audited or safely rolled back.

12.8 2026-Q4 to 2029 watchlist: evidence that would materially change the reading

INT WELL SUPPORTED - Watch Rel-21 WIDs and subsequent RAN3 summaries for any change to the 5G HLS baseline or protocol distribution. A reopening would materially alter the evolutionary-architecture thesis.

INT WELL SUPPORTED - Watch whether RAN3 and SA2 converge on stable data, model and interface assumptions. Common lifecycle or telemetry contracts would materially increase confidence in multi-vendor AI-native interoperability.

INT WELL SUPPORTED - Watch TR 29.832 for adopted protocol primitives, explicit external-protocol reuse and mobile-specific authorization or mobility behavior. The key question is where 3GPP draws the boundary.

INT WELL SUPPORTED - Watch RAN5 for broader AI/ML cases and eventual certification or operator-acceptance use. Multi-vendor reproducibility would move AI from feature proof toward procurement-grade proof.

INT WELL SUPPORTED - Watch power/TCO and live-network scale across operators: watts, cooling, utilization, capex and SLA isolation. Those data can strengthen or weaken the accelerated-common-platform case faster than new PoCs can.

INT WELL SUPPORTED - Watch fleet-scale deployment on existing compute and persistence of KPI gains across bands, topologies and software versions. This would strengthen the software-first path.

INT WELL SUPPORTED - Watch for open telemetry, scoring and policy interfaces plus cross-vendor workload movement under carrier constraints. Portability at this layer would reduce orchestration lock-in risk.

INT WELL SUPPORTED - Watch for named paying customers, repeat use cases, workload density and billing. These would convert a strong narrative into a commercial category.

INT WELL SUPPORTED - Watch Rel-20/21 measurement, architecture and exposure choices together with field accuracy and privacy/governance results. The combination determines whether sensing becomes a reusable service capability.

INT WELL SUPPORTED - Watch comparable energy-per-bit, inference power, site cooling and peak-load data. Energy can become the decisive commercial filter even when performance improves.

INT WELL SUPPORTED - Watch filings that connect RAN control with accelerator-aware scheduling, model lifecycle or agent services. That would test the C20004008 thesis that complementary control points may converge at boundaries.

INT WELL SUPPORTED - Watch explicit liaison or interoperable profiles across 3GPP, O-RAN, IETF and IEEE around agents, data and edge interfaces. Vocabulary alignment is useful; operational interoperability is the stronger signal.

INT WELL SUPPORTED - The watchlist should privilege events that move a conclusion: Rel-21 scope text that changes the HLS baseline; stable lifecycle or data interfaces; reusable conformance assumptions; multi-vendor reproducibility; multi-site power/TCO evidence; repeat external-AI customers; or security/governance rules that materially constrain agent autonomy. Routine announcements are useful context, but they should not move the analytical conclusion unless they change one of those control surfaces.

> **Forecast discipline**
> A watch item matters only if the observable result can strengthen, weaken or falsify a current interpretation. The report deliberately avoids assigning probabilities or ranking standards outcomes. The purpose of the conditional layer is to make later updates auditable: a changed conclusion should point to the new evidence that changed it.

13 METHOD, EVIDENCE BOUNDARY AND CLOSING CLAIMS

13.1 Evidence hierarchy and status language

SRC ESTABLISHED - The evidence cutoff is 18 September 2026. Primary 3GPP meeting records, specification/work-item portals and approved change records anchor formal status claims. ITU-R and O-RAN primary records are used where they provide adjacent framework context. Vendor/operator releases establish attributed implementation evidence. Trade press is used only for attributed commercial-position context that is not available in a primary release.

INT WELL SUPPORTED - Status verbs are deliberately narrow. A proposal is not agreement; a study is not a work item; a work item is not a completed specification; a WG-agreed contribution is not automatically a plenary-approved requirement. Likewise, a live-network trial is not fleet deployment and a PoC is not repeat revenue. These distinctions are central to the report's credibility because AI-RAN activity is moving quickly enough that loose wording can overstate maturity.

13.2 What is not normalized

INT WELL SUPPORTED - Vendor/operator KPI results are not converted into a ranking because public releases do not expose identical topology, band, traffic, model, device, software and power conditions. Alliance membership and demonstration counts are not used as market-share measures. Contribution volume is not used as a vote. The report prefers a narrower conclusion supported by comparable evidence over a broader conclusion created by normalization assumptions.

13.3 Re-derivation triggers

INT WELL SUPPORTED - Re-derive the report after a material Rel-21 scope decision, a change to the 6G RAN HLS baseline, a stable AI lifecycle or data interface, materially new conformance assumptions, a security/governance requirement that changes agent authority, or multi-site operating economics that alter the compute-placement case. A changed record should produce a changed conclusion; stylistic preference should not.

13.4 What this report does not claim

INT WELL SUPPORTED - The report does not claim that 6G requires GPUs, that one AI-RAN architecture has won, that alliance demonstrations equal deployment consensus, that current NR AI/ML conformance machinery is already a 6G regime, or that external AI workloads have become a repeatable operator business. Those claims outrun the public record as of the cutoff.

INT WELL SUPPORTED - It does claim that intelligence is becoming a more durable systems concern across RAN, SA, CT and assurance; that the 6G RAN higher-layer structural path has narrowed; that verification is moving closer to architecture and procurement; and that compute implementation remains commercially plural. Those conclusions can be revisited independently because the report separates the underlying records from the analytical transition.

13.5 Decision checklist

For **standards**, ask which assumption is becoming difficult to reverse and which interface or state definition independent implementations will have to share. For **product**, ask which model, data and telemetry dependencies must remain portable. For **operators**, separate radio KPI value from shared-infrastructure economics and external-AI demand. For **procurement**, ask what can be independently tested, replaced and rolled back. For **policy**, ask who controls the model, operational data, update path and authority to change network state.

Reading the orange ladder

The three levels are **not** a formal 3GPP maturity model. They are a decision aid ordered by how much learned state must become visible beyond the implementation. The gradient is directional: light orange means AI can stay mostly internal; dark orange means more of the surrounding system has to understand, constrain and govern it.

AI-assisted *- Learned optimization stays inside a deterministic envelope. Examples include scheduler or link adaptation, traffic prediction and energy optimization. Commercial proof is persistent KPI gain, acceptable power cost and safe fallback under distribution shift.*

AI-aware *- Model/data state, applicability, telemetry, feedback and fallback become visible to interfaces and assurance. Examples include RAN5 applicability/PICS/test, RAN3 AI/ML framework work and SA6 lifecycle/inference-data handling. Proof becomes portability and repeatable test.*

More deeply AI-native *- Agentic or intent-driven functions can translate goals into network action under identity, authorization, policy, audit and rollback. Examples include the CT3 AI-protocol study, OAM agents/intent and network-assisted Physical AI. Proof becomes bounded autonomy plus safe operational control.*

Standards trend by level

AI-assisted. The standards burden can remain feature-specific: applicability, configuration, performance feedback and fallback around a learned radio function. Current RAN AI/ML work and RAN5 assurance provide the clearest examples.

AI-aware. The burden moves into lifecycle contracts: model/data state, telemetry, versioning, management, service context and reusable assurance. RAN3, SA5, SA6 and RAN5 increasingly touch different pieces of this layer.

More deeply AI-native. Once AI can interpret intent or initiate network action, identity, authorization, policy, audit, rollback and failure containment become protocol and operations requirements. CT3 TR 29.832 and OAM agent/intent work are leading indicators, not proof that one agent architecture is agreed.

Industry and commercial trend: plurality, not one winning compute path

CTX WELL SUPPORTED - Nokia/NVIDIA provide the most visible accelerator-centric path: richer common infrastructure can combine RAN processing with AI workloads if utilization, power, cooling, isolation and orchestration economics hold. Ericsson provides a counterexample with AI-in-RAN functions on existing RAN compute; Samsung explicitly separates AI capability from one mandatory GPU choice. SoftBank explores several paths at once through AITRAS, Nokia external-workload work, Ericsson interworking and Physical-AI programs [14]-[21],[25],[30].

INT WELL SUPPORTED - The strategic signal is therefore not that one camp has won. It is that **standards convergence does not imply compute-platform convergence**. Operators can seek common lifecycle, telemetry and assurance contracts above differentiated silicon and runtimes below.

Commercial proof changes with the level

AI-assisted / AI-for-RAN: the buyer is already the operator and the KPI is already in the network budget. The decisive evidence is persistent spectral-efficiency, throughput, energy or automation gain across bands, traffic, mobility and software versions.

AI-aware / AI-and-RAN: the question becomes infrastructure economics. Shared compute must preserve deterministic RAN behavior while achieving enough utilization to justify capex, power, cooling, transport and orchestration. Multi-site TCO matters more than another PoC.

More deeply AI-native / AI-on-RAN and agents: technical capability is not demand. External workloads need named repeat customers, SLA terms, billing and margin; autonomous network actions need bounded authority and safe rollback. A pilot proves feasibility; repeat operation closes a different risk.

Operator reality check

T-Mobile is informative because it participates across Nokia/NVIDIA and Ericsson initiatives; SoftBank is similarly informative because it spans common infrastructure, orchestration, live-network AI-in-RAN and Physical AI. Multi-path experimentation can be a rational portfolio strategy: one architecture for high-value edge sites, another for broad macro coverage, and different implementations behind a shared lifecycle or orchestration layer.

Verification is market structure

If only the incumbent can reproduce model behavior because data, state or evaluation assumptions are hidden, switching cost rises even when network interfaces are formally open. Shared assurance semantics can preserve proprietary models while making the evidence needed to trust, compare and replace them portable.

Verticals: where higher nativeness can have a reason to exist

Manufacturing and robotics. Locality, deterministic behavior, computer vision and machine control can justify edge inference and network-aware AI. The stronger AI-native case appears when the network contributes sensing, context or policy rather than merely transport.

Automotive, V2X, drones and low-altitude systems. Mobility, positioning, cooperative perception, sensing and uplink media align naturally with AI-assisted radio and ISAC. Deeper nativeness becomes relevant if network-provided context affects decisions; reliability, calibration and liability then become part of the proof burden.

Ports, utilities, private networks and remote systems. These sites can justify richer local compute when sovereignty, resilience, uplink demand or operational autonomy matters. They may be earlier targets for integrated AI/network platforms than uniform macro rollout because value per site and workload predictability are higher.

NTN and constrained links. Intermittent connectivity, constrained energy and long-lived devices make selective data reduction, onboard inference, AI-assisted scheduling and sensing relevant even when inference remains on the device or satellite. The common thread is not "edge AI" as a label; it is that network state materially changes the application outcome.

Conditional outlook: what would move each level

AI-assisted strengthens if fleet-scale KPI and power gains persist across representative networks; it weakens if gains remain narrow, fragile or too expensive to maintain.

AI-aware strengthens if reusable model/data-state, telemetry, fallback and conformance semantics appear across features and vendors; it weakens if proof remains runtime-specific and lifecycle state cannot move between implementations.

More deeply AI-native strengthens if standardized authorization, policy, audit and rollback allow agents or intent-driven functions to act across multi-vendor systems; it weakens if autonomy remains vendor-local or operators cannot bound failure modes.

2027 is the conversion year

The next test is which Rel-20 studies and agreements become concrete enough for Rel-21 normative scope. The March 2027 package/Stage-1 decisions begin that conversion; Stage-2 and Stage-3 freezes in 2028 and coding/API freeze in 2029 progressively reduce room for architectural ambiguity. The high-value signal is therefore not another AI-native label, but a durable interface, lifecycle rule, test procedure, multi-site operating result or repeat customer contract.

Closing synthesis

Structurally familiar does not mean operationally unchanged. A 6G system can preserve recognizable higher-layer splits while making model state, data conditions, assurance and authority increasingly consequential. The practical question is how much of that learned state must cross an interoperability boundary - and whether the evidence needed to trust, replace and govern it can travel with the function.

Practical field guide: evidence that would move the level

AI-assisted - evidence to watch. Fleet expansion of scheduling/link-adaptation and other AI-for-RAN functions; persistent KPI gains across band, topology and traffic; energy delta; fallback under distribution shift; operator acceptance that does not depend on one favored dataset or lab topology.

AI-aware - evidence to watch. Reusable model/data-state semantics, telemetry, applicability, PICS/test assumptions and version/fallback handling across more than one feature; multi-vendor reproducibility; procurement language that can compare or replace learned implementations without exposing every proprietary model detail.

More deeply AI-native - evidence to watch. Stable identity, discovery, authorization, policy, audit and rollback for agents or intent-driven functions; safe failure containment; operational proof that network actions remain bounded across vendors and lifecycle updates rather than being vendor-local automation.

APPENDIX A REFERENCE LEDGER - PRIMARY RECORDS FIRST

Primary standards records precede implementation and market context. Each entry states what it is used for and links to the public record checked for this edition. A linked record is not treated as support beyond its scope: a draft study remains a draft study; an operator trial remains an attributed result; and the prior Apex Standards brief is a historical analytical baseline rather than a standards authority.

[1] Apex Standards C20004008, *The NVIDIA-Nokia AI-RAN Signal: Patent Gravity, 3GPP Intent, IEEE 802.11 Offload, and IETF Agent Protocols*, 24 May 2026. Prior analytical baseline: market clock versus evidence clock; unresolved commercial proof around cost, energy, latency and multi-vendor operation. Source URL

> **Use/boundary:** establishes the report's May 2026 analytical starting point and watch signals. It is cited for continuity of Apex analysis, not as independent proof of 3GPP status.

Re-derivation trigger: compare later evidence against the May watch signals rather than treating the earlier interpretation as fixed. Stronger multi-site economics, changed patent overlap, stable AI lifecycle interfaces or weaker field performance would each change a different part of the original reading.

[2] 3GPP, *Working Group Reports to RAN Plenary #113*, 14 September 2026. Principal source for the Madrid RAN3 HLS agreement, 6G study topics, AI/ML status and ISAC progress. Source URL

> **Use/boundary:** strongest single public summary of the Madrid RAN#113 working-group results used here. It supports reported agreements and study topics, but not broader commercial conclusions.

Re-derivation trigger: a later RAN plenary can narrow or reopen the HLS, AI/ML, ISAC or 6G-study conclusions summarized here. When that happens, the official meeting result supersedes this September snapshot; vendor material does not supersede a plenary agreement.

[3] 3GPP Portal, **TR 38.914**, *Study on 6G Scenarios and requirements*. Technical Report; status Under change control; RAN#112 version 1.0.0 / RP-261565; portal history records the June 2026 change-control transition. Source URL

> **Use/boundary:** anchors document type, title and change-control status for the 6G scenarios/requirements study. It does not imply that all later 6G architecture choices are fixed.

Re-derivation trigger: later approved changes to TR 38.914 can alter scenario wording or requirements context. Because the report uses the document to mark a transition in work character, a future revision should distinguish editorial evolution from a substantive reopening of the scenario text.

[4] 3GPP Portal, WI 1111099, *NR_AIML_NGRAN_Ph3-Core*. Used for NG-RAN AI/ML Phase-3 work-item status and September progress. Source URL

> **Use/boundary:** anchors the formal NG-RAN AI/ML Phase-3 work-item progress used in the longitudinal standards reading. Completion percentage is not treated as deployment maturity.

Re-derivation trigger: completion updates, new CRs or scope changes can move Phase-3 from framework work toward stable interfaces and procedures. The report should change only when those records alter what independent implementations must exchange or support, not merely when percentage completion advances.

[5] 3GPP Portal, NR AI/ML UE-conformance work-item / change-request collection. Used for current-generation applicability, PICS and test machinery; not characterized as a completed 6G regime. Source URL

> **Use/boundary:** shows how current-generation AI/ML features enter familiar conformance artifacts. The report uses this as institutional precedent, not as evidence that 6G conformance is complete.

Re-derivation trigger: broader applicability, reusable test assumptions or cross-feature PICS/data-state requirements would strengthen the claim that AI assurance is becoming a common telecom layer. A collection of isolated feature tests would keep the conclusion narrower and more precedent-oriented.

[6] 3GPP Portal, **TR 29.832**, *Study on the Protocol for Artificial Intelligence in 6G*. Draft Rel-20 Technical Report; created 16 June 2026; v0.2.0 uploaded 14 September 2026. Source URL

> **Use/boundary:** establishes a dedicated 3GPP study home for protocol questions around AI in 6G. Draft status is preserved; no completed standardized agent protocol is inferred.

Re-derivation trigger: adopted identity, authorization, discovery, profile or message semantics would strengthen the protocol-interoperability thesis. Continued dependence on external/proprietary agent frameworks without a mobile-specific contract would weaken the claim that 3GPP is defining a durable AI protocol seam.

[7] 3GPP Portal, **TR 32.801-01**, *Study on 6G Management and Orchestration*. Draft Rel-20 Technical Report; used for management-agent, lifecycle and orchestration context. Source URL

> **Use/boundary:** supports 6G management/orchestration discussion around lifecycle, observability, agents and policy. Draft study material is distinguished from normative deployment obligations.

Re-derivation trigger: stable management-agent registration, intent, observability, lifecycle or rollback semantics would move the analysis from broad OAM direction toward portable operations. If the study retains only high-level use cases, most operational portability will remain implementation-specific.

[8] 3GPP Portal, **TR 33.801-01**, *Study on Security for the 6G system*. Draft Rel-20 Technical Report; used for 6G security context. Source URL

> **Use/boundary:** provides the 6G security-study frame for AI-native operation, identity, authorization and governance concerns. It is not used to claim a finalized security architecture.

Re-derivation trigger: explicit security requirements for agent authority, model/data provenance, lifecycle update or rollback would materially constrain AI-native design. If those questions remain entirely deployment-specific, the report should continue to describe governance as an open dependency rather than a settled contract.

[9] 3GPP SA2#176 TDocsByAgenda, Prague, August 2026. Used for the diversity of 6G architecture, AI-aware service and emerging-service solution work; contribution diversity is not treated as consensus. Source URL

> **Use/boundary:** demonstrates diversity of SA2 architecture proposals and emerging-service solutions during the study phase. Contribution diversity is analytically informative but is not counted as consensus.

Re-derivation trigger: SA2 agreement on a smaller set of architectural solutions matters more than additional contribution diversity. The key change would be a stable system function, reference point or information flow that other groups can rely on; raw TDoc volume is not used as a proxy for support.

[10] 3GPP Portal, AIML_Ph2-APP change-request collection. Used for application-facing inference/model lifecycle and performance-feedback context. Source URL

> **Use/boundary:** supports application-facing inference and lifecycle discussion. Individual CR scope should be checked before using it to support a specific API or feedback requirement.

Re-derivation trigger: individually verified CRs that stabilize inference APIs, collaborative inference, feedback or data requirements would support more specific application-lifecycle claims. Generic collection pages remain insufficient for sentence-level normative assertions unless the relevant CR text has been checked directly.

[11] European Commission 5G Observatory, *Global 3GPP approves Release 21 timeline setting first 6G standards*, June 2026. Used for the explicit March 2027, June 2028, December 2028 and March 2029 milestone schedule. Source URL

> **Use/boundary:** supplies the explicit approved Rel-21 milestone sequence used for the 2027-2029 horizon. The report's description of 2027 as a conversion point is interpretation, not an official 3GPP term.

Re-derivation trigger: 3GPP can revise the milestone sequence, but commercial proof remains independent of the standards calendar. Later editions should distinguish schedule slippage from technical de-scoping, and should not interpret a freeze date as evidence that operator deployment economics have matured.

[12] Recommendation ITU-R M.2160-0, *Framework and overall objectives of the future development of IMT for 2030 and beyond*. Used for formal IMT-2030 Artificial Intelligence and Communication and Integrated Sensing and Communication scenario context. Source URL

> **Use/boundary:** anchors AI and Communication and ISAC as formal IMT-2030 usage-scenario context. It does not select implementation architecture or operator business models.

Re-derivation trigger: later ITU-R evaluation requirements can sharpen how AIAC or ISAC performance is assessed. The scenario framework itself does not decide RAN split, compute substrate or application monetization; those conclusions must continue to come from 3GPP and field evidence respectively.

[13] AI-RAN Alliance, *AI-RAN Alliance Reaches Major Milestone, Showcasing Breakthrough Momentum for AI-Native Networks*, 26 February 2026. Snapshot source for 132 members, 33 MWC 2026 demonstrations and four Industry Blueprints. Source URL

> **Use/boundary:** provides an ecosystem snapshot dated 26 February 2026. Membership, demonstration and blueprint counts indicate breadth of activity, not September market share or deployment consensus.

Re-derivation trigger: later Alliance membership or demonstration counts can update ecosystem breadth but should not move the report's technical conclusion unless blueprints, interfaces or repeat deployments materially change. The snapshot remains useful historically because it shows how wide the ecosystem had become before the September plenary.

[14] Nokia, *Nokia accelerates AI-RAN adoption as global operators embrace AI-native network evolution on NVIDIA platforms*, 16 September 2026. Operator PoC/live-trial breadth and vendor-reported spectral-efficiency evidence. Source URL

> **Use/boundary:** supports Nokia's accelerator-centric implementation path, operator-trial breadth and attributed KPI claims. Reported performance remains vendor context and is not normalized against other vendors.

Re-derivation trigger: multi-site measurements of utilization, power, cooling, isolation and cost are more decision-relevant than additional platform announcements. If accelerated common infrastructure shows persistent economics across ordinary sites, the selective-deployment outlook would need to broaden; weak economics would narrow it.

[15] Ericsson, *The RAN gets smarter: Ericsson puts AI where it matters*, 11 June 2026. Software-centric AI-in-RAN path on Ericsson RAN compute; vendor-reported deployment/trial context. Source URL

> **Use/boundary:** supports Ericsson's software-centric AI-in-RAN path on existing Ericsson RAN compute. It is used to demonstrate implementation plurality, not to rank vendors.

Re-derivation trigger: fleet-scale results showing durable AI gains on installed RAN compute would strengthen the software-first path; evidence that richer acceleration is required for competitive performance would weaken it. The report treats this as an empirical architecture question, not a vendor preference contest.

[16] Ericsson/T-Mobile, *T-Mobile 5G Advanced network achieves world first with Ericsson AI RAN innovation*, 12 May 2026. Live-network attributed result: close to 10% spectral-efficiency improvement and up to 15% downlink-throughput improvement. Source URL

> **Use/boundary:** supports the cited T-Mobile/Ericsson live-network KPI results. The evidence is stronger than a lab demo but still does not establish fleet-wide persistence, power delta or full economics.

Re-derivation trigger: persistence across bands, mobility, traffic mixes and software versions, together with measured power/lifecycle overhead, would convert the trial from a performance signal into stronger product economics. The headline KPI alone is not sufficient for that transition.

[17] SoftBank/Ericsson Japan, *SoftBank Corp. and Ericsson Conduct Japan's First Trial of Ericsson AI in RAN on a 5G Commercial Network*, 20 August 2026. Attributed site and average KPI results. Source URL

> **Use/boundary:** provides an additional live commercial-network data point from Japan and shows site-specific versus average gains. Conditions differ from other trials and are therefore not cross-normalized.

Re-derivation trigger: repeated results across a broader Japanese footprint or independent test conditions would strengthen generalization. If average gains fall materially outside the initial locations, the evidence should remain site-specific rather than being generalized into a network-wide expectation.

[18] SoftBank, *From Proof-of-Concept to Commercial Deployment of AI-RAN*, 16 July 2026. Operator R&D;/commercialization context for AITRAS. Source URL

> **Use/boundary:** captures SoftBank's operator R&D; and commercialization narrative around AITRAS. It informs strategy context but does not substitute for audited commercial revenue or third-party economics.

Re-derivation trigger: named commercial deployments, utilization data or operator financial disclosure would move AITRAS from commercialization narrative toward measurable business evidence. Until then, the source remains valuable for understanding operator intent and engineering direction rather than realized margin.

[19] SoftBank/Nokia, *SoftBank Corp. and Nokia to Enable Execution of External AI Workloads on AI-RAN*, 25 February 2026. External-workload and new-revenue concept; date follows the individual primary release. Source URL

> **Use/boundary:** supports the external-AI-workload proposition and its February 25 primary-release date. It demonstrates capability direction, not repeat customer demand, billing or margin.

Re-derivation trigger: recurring external workloads, SLA commitments, billing and customer retention are the missing facts. A larger set of technically successful offload demonstrations without paying repeat users would not close the AI-on-RAN business case identified in the body.

[20a] SoftBank/Ericsson, *AITRAS Orchestrator and Ericsson Intelligent Automation Platform interworking*, 27 February 2026. Dynamic allocation/orchestration context. Source URL

> **Use/boundary:** supports orchestration interworking between SoftBank AITRAS and Ericsson automation. It is evidence of control-plane integration, not proof of multi-vendor workload economics at scale.

Re-derivation trigger: portable orchestration across additional RAN/AI platforms, with common telemetry and failure handling, would strengthen the interoperability thesis. One successful interworking path shows feasibility but not yet the substitutability or operating simplicity required for broad procurement confidence.

[20b] SoftBank/Ericsson, *Network-enabled Physical AI With AI-RAN*, 27 February 2026. Physical-AI PoC using MEC offload and differentiated connectivity. Source URL

> **Use/boundary:** supports a network-enabled Physical-AI PoC. It demonstrates technical integration of connectivity, MEC and application behavior but not safety assurance or repeatable revenue.

Re-derivation trigger: repeat Physical-AI deployments with quantified reliability, safety controls and application economics would elevate this from PoC evidence. If the capability remains bespoke to individual sites or demonstrations, it should continue to be treated as an application exploration rather than a market category.

[21] NVIDIA/T-Mobile/partners, *Integrate Physical AI Applications on AI-RAN-Ready Infrastructure*, 16 March 2026. Ecosystem Physical-AI context. Source URL

> **Use/boundary:** provides broader Physical-AI ecosystem context around T-Mobile/NVIDIA and partners. It is treated as implementation/ecosystem evidence rather than a standards record.

Re-derivation trigger: named enterprise users and repeat workloads operating on the infrastructure matter more than expansion of the partner list. The report will strengthen the AI-service thesis only when application demand, workload density and service economics become observable rather than assumed.

[22] Keysight/Samsung/NVIDIA, *End-to-End AI-RAN Validation Workflow*, 2 March 2026. Ecosystem demonstration linking data generation, model training and benchmarking. Source URL

> **Use/boundary:** demonstrates an end-to-end validation workflow linking scenarios/data, model training and benchmarking. It illustrates assurance direction; it is not an industry-wide conformance regime.

Re-derivation trigger: reproducible results across different tools, models and vendor implementations would show that the validation workflow is becoming portable assurance infrastructure. A workflow that depends on one tightly coupled ecosystem remains useful but does not establish an open test contract.

[23] O-RAN Alliance, *Digital Twin RAN Use Cases*. External-SDO context on AI/ML training/evaluation/testing, test automation, planning, energy saving and site-specific optimization. Source URL

> **Use/boundary:** supports the digital-twin use-case taxonomy outside 3GPP. The report uses it to explain repeatable evidence environments while retaining the need to validate the twin itself.

Re-derivation trigger: evidence that digital-twin predictions remain accurate under field change, mobility, traffic and hardware variation would strengthen their role as procurement-grade evidence. Poor sim-to-real correlation would limit twins to planning/training assistance rather than high-confidence acceptance testing.

[24] Fierce Network, attributed operator interviews on divergent U.S. AI-RAN/GPU positions, 18 September 2026. Secondary commercial-position context only. Source URL

> **Use/boundary:** secondary source used only for attributed comparison of U.S. operator positions where a single primary comparative record is unavailable. It is not used for formal standards status.

Re-derivation trigger: direct operator technical or financial records should replace this secondary context when available. The cited article is retained only to show explicitly attributed divergence among operators; it should never carry a factual claim that can be grounded in a primary company source instead.

[25] Samsung, *Samsung at MWC 2026: Adaptive, AI and Software-Driven Networks in Action*, 7 April 2026. Explicit CPU/GPU plurality and statement that AI does not equal GPU. Source URL

> **Use/boundary:** directly supports Samsung's statement that AI does not equal GPU and that compute choice depends on deployment economics and use case. It reinforces compute-path plurality.

Re-derivation trigger: later Samsung deployment evidence can show where CPU, GPU or other accelerators are actually selected in products. The article's strategic value is the explicit separation of AI capability from one hardware category, not a forecast that any specific compute mix will dominate.

[26] 3GPP Portal, **TR 23.801-01**, *Study on Architecture for 6G System; Stage 2*. Draft Rel-20 Technical Report, SA2. Source URL

> **Use/boundary:** anchors the formal SA2 6G Stage-2 architecture study. It provides system-level context without implying that contribution-level options inside the study are agreed architecture.

Re-derivation trigger: agreed Stage-2 functions, interfaces and information flows will narrow which AI/service-management concepts become architectural contracts. Until that point, solution proposals inside the study remain options; the report should not collapse them into one 6G core design.

[27] 3GPP RAN#113 records, Artificial Intelligence/Machine Learning for NR air-interface enhancements, status material referenced in the official RAN#113 working-group report. Source URL

> **Use/boundary:** supports the continuing RAN air-interface AI/ML work summarized at Madrid. Study/work-item status is kept distinct from commercial-product or conformance maturity.

Re-derivation trigger: approved work scope, stable evaluation methodology or normative hooks would strengthen the air-interface AI/ML path. Continued study without interoperable procedures would keep most implementation value proprietary and limit what RAN5 can later verify across vendors.

[28] 3GPP Portal, RAN5#112 / RP-262069-related AI/ML change-request records in 38.523-1 and associated specifications. Used for current-generation test/data-collection evidence and as procedural precedent. Source URL

> **Use/boundary:** supports the RAN5 test/data-collection discussion and the procedural precedent for AI/ML assurance. Exact CR scope governs any feature-specific conclusion.

Re-derivation trigger: common applicability/state descriptions across multiple AI/ML features would be especially significant because they could become reusable procurement language. If each feature requires bespoke evidence and hidden model assumptions, assurance will remain fragmented and switching cost higher.

[29] 3GPP RAN#113/RAN3, NR sensing / ISAC progress in the Madrid working-group report; used for standards trajectory rather than vertical-market proof. Source URL

> **Use/boundary:** anchors the 3GPP sensing/ISAC standards trajectory used in the vertical discussion. It is not evidence that any particular automotive, city or industrial revenue model is proven.

Re-derivation trigger: stable sensing measurements, reporting architecture and privacy/security controls would make vertical use cases easier to translate into deployable services. If field calibration or governance proves highly application-specific, ISAC may remain a capability with narrower commercial reuse.

[30] SoftBank, *Open-sources AITRAS Orchestrator to Expand AI-RAN Ecosystem*, 18 February 2026. Open orchestration context. Source URL

> **Use/boundary:** supports open-orchestration context for AITRAS and is used when discussing portability and ecosystem direction. Open-source availability alone does not prove interoperability or economics.

Re-derivation trigger: deployment by independent ecosystems and telemetry/policy portability would show that open orchestration is reducing switching friction. Open-source code without interoperable operational contracts can still reproduce proprietary coupling at the model, runtime or infrastructure layer.

Editorial boundary

Market visibility does not become technical consensus; contribution volume does not become agreement; an alliance blueprint does not become a 3GPP requirement; and a proof of concept does not become repeatable revenue. The report's purpose is to show how those evidence classes interact without collapsing them.

About APEX STANDARDS

Apex Standards provides high-efficiency analytical tools and expert advisory services for global standardisation and patenting. We offer advanced software platforms that analyse complex data across 3GPP, IEEE, IETF, ITU, Open RAN, video codec, JEDEC, AI infrastructure, CXL, HBM and related technology ecosystems. We equip governments, research institutes, and corporate innovators with the rigorous standards and intellectual property intelligence needed to drive R&D excellence.

Website: www.apexstandards.com

Email: support@apexstandards.com