

Large Wireless Models: A Survey of Foundation Models for 6G Communications and Sensing

Jiseok Youn

Department of Electrical and Computer Engineering and INMC
Seoul National University
Seoul, South Korea
jsyoun@netlab.snu.ac.kr

Saewoong Bahk

Electronics and Telecommunications Research Institute
Daejeon, South Korea
sbahk@etri.re.kr

Abstract—Foundation models (FMs) have reshaped language and vision by replacing task-specific pipelines with one self-supervised model that transfers across downstream tasks. Wireless communications now face the same pressures—costly labeled data, poor generalization across deployments, and a proliferation of per-task models—which have crystallized into a direction often termed the Large Wireless Model (LWM). We survey this landscape under a deliberately inclusive definition: a large-scale, task-agnostic pretrained model serving many networking, communication, and sensing tasks under minimal task-specific adaptation. Since the definition does not restrict input modality, wireless-native FMs and text-based telecom large language models are both admitted, separated by design path rather than by membership. We organize recent work into physical-layer, network-traffic, generative/telecom-LLM, and multimodal/ISAC categories, record design path and pretraining objective as cross-cutting attributes, and identify the limitations blocking deployment. We give particular weight to evaluation integrity—shortcut exploitation, pretraining–benchmark overlap, and comparison at uncontrolled adaptation budgets—which the traffic category has documented and developed cheap remedies for, whereas the physical-layer line, pretraining and evaluating on the same small set of synthetic scenario libraries, has not been audited the same way.

Index Terms—Foundation models, large wireless models, 6G, self-supervised learning, channel state information, ISAC, AI-RAN.

I. INTRODUCTION

The integration of artificial intelligence (AI) into the radio access network is a defining theme of the evolution toward 6G. The 3GPP study on AI/ML for the NR air interface examined channel state information (CSI) feedback, beam management, and positioning, recommending normative work for the latter two [1], while edge intelligence has become a core enabler of 6G, with AI treated as a native design tool rather than an optimization add-on [2].

Yet the dominant pattern remains *task-specific*: one model per function, band, and deployment. Supervised training needs labeled data that is expensive to collect in wireless [3], models tuned to one environment degrade out of distribution [4], and the proliferation is hard to validate and operate—one CSI-feedback study trained forty models to cover a single grid of configurations [5]. A foundation model (FM) is trained on broad data, generally by self-supervision at scale, then adapted to many downstream tasks [6]; reviews of the wireless

case add that capability strengthens with parameters, data, and compute, and that the recipe already transferred to domains such as biomedicine and weather forecasting [13]. The premise is not that radio data rivals web text—the early case for a physical-layer FM argued the opposite [7]—but that one model pretrained without labels can be reused across tasks for which task-specific labelled data is scarce [8].

A. Defining Large Wireless Models

We use *Large Wireless Model* (LWM) for a large-scale, task-agnostically pretrained model supporting many downstream networking, communication, and sensing tasks under *minimal task-specific adaptation*, ranging from zero-shot inference, through few-shot inference and lightweight heads on frozen representations, to parameter-efficient fine-tuning.

Two aspects are deliberate. Adaptation cost is a spectrum, not a membership test: zero-shot inference is the ideal endpoint and a few models reach it [5], [17], but most still attach a small head to a frozen or lightly tuned backbone [3], [9], [18], so a strict zero-shot criterion would exclude the most mature line of work. And we do not restrict input modality: a model over textual telecom artifacts is an LWM if it is large, pretrained, and serves many networking tasks cheaply, what separates it from a wireless-native model being *design path*. The word *large* also deserves scrutiny, since backbones span 0.6 M to a few hundred million parameters with no monotone relation to downstream breadth; following [13] we read it as capability—coverage of heterogeneous configurations, breadth of tasks, cross-scenario generalization—rather than parameter count. Overviews go further, defining a wireless FM not by size but by adaptation across tasks, scenarios, and system configurations plus non-trivial transferability *beyond the pre-training distribution* [8], which Section III argues is what current evaluation fails to test.

Section II groups recent work into (i) physical-layer channel and signal FMs, (ii) network-traffic and management FMs, (iii) generative and telecom-LLM approaches, and (iv) multimodal and integrated sensing and communication (ISAC) FMs. Categories (i), (ii), and (iv) are organized by *domain*, whereas (iii) is organized by *design path*: whether a model is pretrained wireless-native or built on a pretrained text large language model (LLM) [10], [12], [26]. This split is formalized as two

paradigms with a selection rule: borrow a pretrained backbone when wireless data can be aligned with its source modality and latency is relaxed; build native when physical-layer timing and device power budgets bind [13]. Design path therefore cuts across the other categories—LLM4CP is both a physical-layer effort and an instance of repurposing—so we record it and the *pretraining objective* as columns of Table I. A third paradigm—agentic models that plan and invoke external solvers rather than encode a signal, which overviews place at the end of this trajectory [8]—is excluded: such systems are orchestrators built *on* foundation models.

Two themes recur. Wireless data is not text: inputs are complex- and matrix-valued and span widely varying sizes, so tokenization, positional encoding, and normalization cannot be borrowed unchanged, and every repurposing effort pays this in bespoke preprocessing and output modules [10], [11], [26]. And scale alone does not resolve it: CSI-reconstruction error falls as a power law in corpus size and parameter count without saturating, but an order of magnitude more of either buys only about a fifth of the error [5], while the deployment envelope is hard [13].

II. RECENT LWM WORK

We group recent LWM research into the four categories of Table I, which also records design path and pretraining objective as cross-cutting attributes.

Three self-supervised objectives recur. *Masked and reconstructive modeling* recovers hidden channel, signal, or packet elements from context and dominates, underlying LWM [3], BERT4MIMO [14], LWM-Spectro [16], 6G WavesFM [19], and—despite its name—WirelessGPT [15]. A variant substitutes an explicit information bottleneck: WiFo-CF compresses CSI to a quantized codeword and reconstructs it jointly across users at randomly sampled feedback rates, so the pretext task is the deployment task [5]. *Contrastive learning* pulls together views sharing a propagation state, using masking and noise as augmentations [9] or aligning a radio measurement with a co-located camera frame [18]. *Autoregressive generation* appears mainly in traffic [23], and alone buys a second capability, since the weights that classify can also synthesize. Two qualifications matter. Masking is rarely used alone above the physical layer, where traffic FMs add objectives over larger structural units—ET-BERT’s same-origin BURST prediction [24], netFound’s burst-level tasks [25]—since a token-level loss cannot teach flow-level coherence. And an entire column of Table I has no pretraining objective: models repurposing a text LLM inherit their prior from language pretraining and add only adapters, so design path and objective are not independent axes.

A. Physical-Layer Channel and Signal FMs

The most mature line treats the channel as the pretraining substrate under a shared template: patches across antenna and subcarrier dimensions are linearly embedded and processed by a transformer encoder, so each token summarizes a local region while self-attention captures global correlation.

TABLE I
REPRESENTATIVE LWM EFFORTS. CATEGORIES 1, 2, AND 4 ARE ORGANIZED BY DOMAIN; CATEGORY 3 BY DESIGN PATH, WHICH CUTS ACROSS THEM.

Model	Design path	Pretraining objective	Adaptation
<i>1) PHY-layer channel and signal FMs</i>			
LWM [3]	Native	Masked (MCM)	Frozen emb. + head; partial FT
LWM-Spectro [16]	Native (MoE)	Masked; contrastive at FT	Frozen or low-LR enc.; few-shot
BERT4MIMO [14]	Native	Masked	Single task
ContraWiMAE [9]	Native	Masked AE + contrastive	k NN probe on frozen enc.
WiFo-CF [5]	Native (MoE)	Multi-user, multi-rate compression	Zero-shot ; frozen + head
Tiny-WiFo [17]	Native (distilled)	Distillation (WiFo teacher)	Zero-shot ; 5.5 M params
<i>2) Network-traffic and management FMs</i>			
NetGPT [23]	Native	Autoregressive (hex tokens)	FT with prompts; header shuffling
ET-BERT [24]	Native	Masked + same-origin BURST	End-to-end FT; few-shot
netFound [25]	Native	Hierarchical masked + 3 burst tasks	Frozen enc. + head; end-to-end FT
NetLLM [26]	Repurposed LLM	None (text-LLM prior)	Multimodal enc., head, LoRA
<i>3) Generative and telecom-LLM approaches</i>			
WirelessGPT [15]	Native	Masked	Frozen rep. + head; optional FT
LLM4CP [10]	Repurposed LLM	None (text-LLM prior)	Partly frozen GPT-2 + task modules
LLM4WM [11]	Repurposed LLM	None (text-LLM prior)	Frozen GPT-2 + MoE-LoRA; 6 tasks
<i>4) Multimodal and ISAC FMs</i>			
WMFM [18]	Native	Cross-modal contrastive	Frozen encoders + light heads
6G WavesFM [19]	Native	Masked	Last 2/12 blocks FT; or frozen + LoRA
Aboufotouh and Abou-Zeid [20]	Native	Masked (multimodal, shared ViT)	Linear probe; 2-block FT; LoRA

LWM [3], widely cited as the first FM for wireless channels, is pretrained by masked channel modeling on over one million channels from fifteen ray-traced DeepMIMO scenarios [21]; its backbone holds only 600 K parameters, an early sign that “large” here means pretraining scale. LWM-Spectro [16] extends this to 9.2 M baseband spectrograms across WiFi, LTE, and 5G, adding three protocol-specialized encoders under a top-1 router so protocol-aware features cost one expert. BERT4MIMO [14] applies the same objective to massive-MIMO CSI reconstruction, but for one task on statistically generated data, and ContraWiMAE [9] contrasts two differently masked, noise-perturbed views of the *same* channel, probing its encoder by k -nearest neighbors to avoid optimization bias. WiFo-CF [5] is the clearest zero-shot instance of our definition: shared and routed experts over multi-user patched CSI, pretrained across heterogeneous antenna counts, subcarrier allocations, and quantization widths, it applies to unseen configurations and there beats most baselines trained to convergence. Tiny-WiFo [17] distills the WiFo channel-prediction model into a 5.5 M-parameter student retaining over 98% of its teacher’s performance and its zero-shot generalization, and LWLM [4] adds domain-transformation invariance and position-invariant contrastive learning for localization.

Pretraining data is predominantly synthetic and of two kinds—ray traced, as in DeepMIMO [21], or statistically generated—a distinction that decides whether environment geometry is represented.

B. Network-Traffic and Management FMs

A second category treats packet and flow traces as a hierarchical “language” with protocol syntax, temporal rhythms, and usage semantics. A systematic review of 51 such studies finds tokenization schemes incompatible across models, cross-domain benchmarks absent, and cross-task transfer limited; its verdict on corpora is sharper than a bare scarcity claim, since two openly licensed 1.2TB backbone corpora exist, so the binding constraint is diversity, not volume [22]. ET-BERT [24], the reference point, pretrains a 110M-parameter transformer on about 30GB of unlabeled traffic segmented into BURSTs and retains about 91.6% F1 with a tenth of the labels where a supervised baseline loses over 40 points—the label-efficiency argument for pretraining, made in traffic rather than at the physical layer. NetGPT [23] is the autoregressive counterpart, its synthesized header fields reaching a Jensen–Shannon divergence of 0.041 against real traces, against roughly 0.1–0.6 for GAN-based generators. netFound [25] is the most thoroughly engineered instance, its architecture derived from a diagnosis of why earlier traffic FMs fail: protocol-aware tokenization splits headers on field boundaries so no token bisects a 16-bit value; operational context such as inter-arrival time is concatenated rather than added, so later layers can separate the two sources; hierarchical attention runs within bursts and then over burst summaries, matching the two time scales at which congestion control and cross-traffic manifest; and payload and endpoint addresses are excluded by construction. It pretrains on over four billion real flows, released in full. NetLLM [26] instead equips a frozen Llama2-7B with per-modality encoders and low-rank adapters, replacing the language-modeling head with a per-task head emitting one answer from the valid output set in a single pass, since token-by-token decoding admits physically invalid answers and took 3.84s against a 1s deadline. Architectural breadth here also extends past the transformer, [22] recording a 2.2 M-parameter state-space backbone with up to $60\times$ higher throughput on long flows. Unlike the physical layer, this category pretrains predominantly on real measured traffic, simulator-generated corpora being the exception rather than the rule [22]—its principal advantage, and what has allowed its evaluation to be audited more sharply.

C. Generative and Telecom-LLM Approaches

A third category is defined by how the model is built and so intersects the others. WirelessGPT [15] unifies channel estimation, channel prediction, human activity recognition, and environment reconstruction; despite the name its objective is masked reconstruction. LLM4CP [10] adapts GPT-2 to channel prediction, freezing most of its first six layers so 1.7M of 82.9M parameters train—a boundary case: a foundation model borrowed without wireless-native pretraining, applied to a

single task. LLM4WM [11] removes that caveat with mixture-of-experts low-rank adapters and per-task heads, serving six channel-associated tasks at about 1.1M parameters each. At the operations layer, telecom LLMs [12] work on standards, configurations, and troubleshooting logs, and their network-modeling experiment cuts both ways: asked to model base-station energy consumption, GPT-3.5 selected the right 5 of 12 inputs without seeing data but summed where load and maximum transmit power should multiply; domain context cut relative error from 7.8% to 3.0%, against 2.3% for an expert design, whereas LLaMA-70B could not be corrected by it—domain context, not backbone scale, closed the gap. Under Section I-A these are LWMs, not a foil to them. The route raises an unsettled representational question: a text LLM emits a distribution over a fixed vocabulary whereas channels are continuous and multi-path structured, so an adapter must both project the signal into the embedding space and replace the output head [11]. Interface quality matters more than borrowed backbone size—eight GPT-2 layers instead of six degrade accuracy while increasing latency [10]—and what a borrowed backbone contributes is the language prior rather than its parameter count, since reinitializing it degrades every task sharply [26].

D. Multimodal and ISAC FMs

Since CSI alone gives a partial view of an environment and is sensitive to noise, blockage, and mobility, multimodal FMs fuse complementary modalities. WMFM [18] aligns vision and communication by a symmetric contrastive objective over paired camera frames and channel coefficients from the ray-traced DeepVerse6G dataset, then freezes both encoders behind lightweight heads; its convolutional encoders show the transformer template is not universal. 6G WavesFM [19] instead shares one Vision Transformer across 5G NR positioning, MIMO-OFDM channel estimation, human activity sensing, and RF signal classification, unfreezing only the last two of twelve blocks and matching that with LoRA on a fully frozen backbone; it is also one of the few models pretrained entirely on real measurements, though on 4648 samples—a counterpoint to the sim-to-real concern of Section III that costs scale. The multimodal wireless FM of Aboufotouh and Abou-Zeid [20] accepts raw IQ alongside image-like modalities under one masked objective, sharing a single 6.3M-parameter ViT and differing only in the input path—patch embedding for image-like inputs, segment embedding for IQ, and a learned per-antenna embedding atop the temporal positional code, since array geometry and propagation relate the streams in a way sinusoidal position cannot express. Because dedicated multimodal FMs are few, progress leans on shared datasets: DeepSense 6G [27] supplies over a million real data points across more than forty deployment scenarios in two countries, and overviews catalogue larger synthetic collections and real in-the-wild benchmarks alongside it [8].

III. LIMITATIONS AND FUTURE OPPORTUNITIES

Current LWMs remain early-stage. We organize their limitations into research opportunities.

A. Wireless-Native Representation, Architecture, and Cost

Two recent designs show what a domain-native answer looks like, and neither is a large model: in the multimodal WFM the input path is the only difference between IQ and image-like modalities, enough for a 6.3 M-parameter shared encoder to beat a dedicated IQ backbone [20], while netFound’s choices answer measured failures of byte-window tokenization and additive fusion [25]. Concrete proposals for radio-native tokenization—byte-pair encoding over co-occurring IQ pairs, pulse-level atomic tokens—predate both on thin evidence and have never been evaluated inside a released model against the patch-and-project template every entry in Table I adopts [7]; that missing comparison is itself the opportunity. Architecture matters because scale cannot substitute for it: base-station distributed units offer on the order of 100 TOPS, handsets under 10 TOPS, and physical-layer functions must complete at the millisecond level at five-nines reliability, so most LWMs sit below 300 M parameters and wireless-native ones below 100 M [13]. Sparse mixture-of-experts is the common answer [5], [16], whereas the traffic category reaches for linear-complexity state-space backbones [22], an option the physical layer has flagged but not realized despite its tighter latency budget [13]. Compression needs equal care: Tiny-WiFo reaches 1.6 ms inference, yet conventional logits-matching distillation of the identical architecture performed *worse* than the same student trained with no teacher [17]. The deployment target decides feasibility: the ViT of 6G WavesFM costs 14.9 GFLOPs per pass, roughly 25 ms on a Jetson Orin Nano [19], far beyond a millisecond budget. For a technology whose promise includes energy efficiency [2], cost must be weighed against amortization: the roughly 23 GPU-hours of one WiFo-CF run replace training forty configuration-specific baselines [5]. Yet across 51 traffic studies almost none report energy at all [22].

B. Pretraining Data and Out-of-Distribution Generalization

Physical-layer LWMs are predominantly pretrained on synthetic channels [21], and the two clearest exceptions pull in opposite directions: WiFo-CF combines statistical models, ray tracing, and real measurements [5], whereas 6G WavesFM pretrains solely on real measurements but on only 4 648 samples [19]. That trade is loosening at both ends, though not through the models surveyed here: overviews catalogue synthetic CSI collections of order 10^{10} points and in-the-wild Wi-Fi CSI benchmarks of hundreds of hours [8], neither yet a pretraining substrate for any model in Table I. Realism is two axes—fidelity per sample and coverage of the deployment population—and netFound’s four billion real flows come from a single campus border [25]. DeepSense 6G is the closest the wireless side has to a corpus strong on both [27], yet serves almost entirely as a supervised benchmark, for a structural reason: real-world campaigns record received signal strength,

beam indices or raw waveforms rather than the decoupled CSI a masked-channel objective consumes [8], which is precisely what the in-the-wild CSI benchmarks noted above supply—turning one into a pretraining substrate is a concrete opportunity. More data is not automatically better: a model pretrained on 5G CSI *underperforms* on 5G NR positioning despite the nominal match, because hardware, environment, and preprocessing differ—negative knowledge transfer [19]. Two notions of shift are likewise conflated, environment and system configuration, and they behave differently: LLM4CP transfers zero-shot from a 3GPP UMa to a UMi scenario, yet the same model carried from 2.4 to 4.9 GHz fails zero-shot, as does every deep baseline tested [10]; ContraWiMAE overtakes a supervised baseline on unseen-scenario beam labels once noise is present [9]; and LWLM transfers to unseen bandwidths yet pretrains on one DeepMIMO scenario [4], whereas LWM spans fifteen scenarios at a fixed array [3]. The traffic category adds a third notion worth importing—*exogenous context*: the congestion-control algorithm and cross-traffic under which an otherwise identical flow was generated, invisible in packet content and manifest only in timing, which a linear probe on netFound separates at F1 0.95 against 0.62 for the best of five published traffic FMs [25]. The physical-layer analogue—scheduler behavior, interference regime, hardware impairment—has no comparable probe, and building one would test representation quality far more sharply than another held-out ray-traced scenario.

C. Evaluation, Standardization, and Trust

A useful benchmark would fix not only datasets and metrics but the adaptation budget, since much of the appeal of LWMs lies in the low-label regime. Table I shows models compared as peers sitting at very different points on that spectrum, so a comparison not holding the budget fixed is not measuring representation quality; WiFo-CF’s headline result places a zero-shot model against baselines trained to convergence, the right comparison for its claim but not a like-for-like one [5]. More corrosive is that some numbers do not measure what they appear to, and two failure modes are documented in traffic with no reason to think the physical layer immune. The first is shortcut exploitation: removing a class-unique timestamp field from one benchmark drops ET-BERT’s top-10 accuracy from 99.8% to 32.8%, while netFound, which never sees the leaking field, is unchanged [25]. The second is overlap between pretraining corpora and benchmarks: pretraining and fine-tuning on random splits of one corpus reduces a foundation model to an unsupervised feature extractor for a single distribution, the remedy being a cross-domain benchmark that pretrains on one deployment and evaluates on an unrelated one [22]. Both are invisible to evaluations reporting only fine-tuned accuracy. Physical-layer LWMs pretrain and evaluate on the same handful of synthetic scenario libraries—structurally the same situation—and the corresponding remedy, pretrain on a scenario library and evaluate on a measurement campaign, is well defined and, as far as we are aware, untried. Frozen-encoder evaluation is the cheapest safeguard, since model ordering

changes between regimes: one traffic FM scores 0.14 frozen and 0.94 fine-tuned on the same task [25]. Reproducibility depends first on what is released—in traffic, roughly a third of models publish code and fewer publish weights [22]—and suffers further from results insensitive to the intervention under test, such as BERT4MIMO’s cross-scenario error matrix being nearly invariant to the training scenario [14]; candor about equivocal ablations is the other half, and NetGPT reports that removing one of its two mechanisms *improves* the flow-level average [23]. For the LLM branch the gap concerns reliability, since GPT-3.5’s error probability in selecting the strongest of N beams under an exclusion constraint rises monotonically with N [12]; overviews extend the charge to work demonstrating one or two scenarios without systematic evaluation under impairments or real-time constraints [8].

Rigorous evaluation is also the bridge to standardization. The 3GPP study leaves CSI compression without a normative recommendation, citing the performance–complexity trade-off and the difficulty of inter-vendor training collaboration for the two-sided encoder–decoder pair; its options range from exchanging a dataset or encoder parameters across the vendor boundary to fully standardizing a reference model, the latter eliminating collaboration complexity at high specification cost [1]. A shared, pretrained channel representation is a natural candidate for that reference interface. Deployment then raises governance, and restricting what enters the model is the cheapest answer—netFound excludes payload and endpoint addresses by construction yet yields the strongest frozen representations, so privacy and representation quality improve together [25]. A shared backbone also concentrates risk, since one poisoned representation propagates to every task reusing it: “toxic” subword embeddings injected into a pretraining corpus can install a backdoor surviving ordinary fine-tuning, and how to construct them for encrypted traffic has not been studied [24]—an unstudied attack is not an absent one.

IV. CONCLUSION

We surveyed the emerging landscape of LWMs under a deliberately inclusive definition and organized recent efforts into physical-layer, network-traffic, generative/telecom-LLM, and multimodal/ISAC categories; because the third is defined by design path rather than by domain, we recorded design path and pretraining objective as cross-cutting attributes, which keeps text-based telecom LLMs inside the definition instead of setting them against wireless-native models. The paradigm promises to replace many label-hungry narrow models with one that generalizes and adapts cheaply, yet the field is early: few models operate genuinely zero-shot, and backbones range from under a million to a few hundred million parameters with no settled account of what makes an LWM “large.” The more useful criterion is transferability beyond the pretraining distribution—exactly what current practice does not test. The most consequential blockers are the scale and realism of pretraining data, the need for wireless-native representations under hard latency limits, alignment with standardization, and the integrity of evaluation.

REFERENCES

- [1] “Study on artificial intelligence (AI)/machine learning (ML) for NR air interface,” 3GPP, Sophia Antipolis, France, Rep. TR 38.843 V19.0.0, Sep. 2025.
- [2] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, “Edge artificial intelligence for 6G: Vision, enabling technologies, and applications,” *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.
- [3] S. Alikhani, G. Charan, and A. Alkhateeb, “LWM: A pre-trained wireless foundation model for universal feature extraction,” in *Proc. IEEE Int. Conf. Mach. Learn. Commun. Netw. (ICMLCN)*, Barcelona, Spain, May 2025, pp. 1–6.
- [4] G. Pan *et al.*, “Large wireless localization model (LWLM): A foundation model for positioning in 6G networks,” 2025, *arXiv:2505.10134*.
- [5] X. Liu *et al.*, “WiFo-CF: Wireless foundation model for CSI feedback,” *IEEE Trans. Wireless Commun.*, vol. 25, pp. 15039–15053, 2026.
- [6] R. Bommasani *et al.*, “On the opportunities and risks of foundation models,” 2021, *arXiv:2108.07258*.
- [7] J. Fontaine, A. Shahid, and E. De Poorter, “Towards a wireless physical-layer foundation model: Challenges and strategies,” in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2024, pp. 1566–1572.
- [8] L. Liang *et al.*, “Foundation models for wireless communications: From PHY intelligence to network autonomy,” 2026, *arXiv:2606.06239*.
- [9] B. Güler, G. Geraci, and H. Jafarkhani, “A multi-task foundation model for wireless channel representation using contrastive and masked autoencoder learning,” *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 4489–4504, 2026.
- [10] B. Liu *et al.*, “LLM4CP: Adapting large language models for channel prediction,” *J. Commun. Inf. Netw.*, vol. 9, no. 2, pp. 113–125, Jun. 2024.
- [11] X. Liu *et al.*, “LLM4WM: Adapting LLM for wireless multi-tasking,” *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, pp. 835–847, 2025.
- [12] A. Maatouk *et al.*, “Large language models for telecom: Forthcoming impact on the industry,” *IEEE Commun. Mag.*, vol. 63, no. 1, pp. 62–68, Jan. 2025.
- [13] X. Cheng, B. Liu, X. Liu, and X. Cai, “Large wireless foundation models: Stronger over bigger,” 2026, *arXiv:2601.10963*.
- [14] F. O. Catak, M. Kuzlu, and U. Cali, “BERT4MIMO: A foundation model using BERT architecture for massive MIMO channel state information prediction,” 2025, *arXiv:2501.01802*.
- [15] T. Yang *et al.*, “WirelessGPT: A generative pre-trained multi-task learning framework for wireless communication,” *IEEE Netw.*, vol. 39, no. 5, pp. 58–65, Sep. 2025.
- [16] N. Kim, S. Alikhani, and A. Alkhateeb, “LWM-Spectro: A foundation model for wireless baseband signal spectrograms,” 2026, *arXiv:2601.08780*.
- [17] H. Zhang, S. Gao, and X. Cheng, “Tiny-WiFo: A lightweight wireless foundation model for channel prediction via multi-component adaptive knowledge distillation,” *IEEE Wireless Commun. Lett.*, vol. 15, pp. 1846–1850, 2026.
- [18] M. Farzanullah *et al.*, “Wireless multimodal foundation model (WFMF): Integrating vision and communication modalities for 6G ISAC systems,” 2025, *arXiv:2512.23897*.
- [19] A. Aboulfotouh, E. Mohammed, and H. Abou-Zeid, “6G WavesFM: A foundation model for sensing, communication, and localization,” *IEEE Open J. Commun. Soc.*, vol. 6, pp. 6792–6807, 2025.
- [20] A. Aboulfotouh and H. Abou-Zeid, “Multimodal wireless foundation models,” 2025, *arXiv:2511.15162*.
- [21] A. Alkhateeb, “DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications,” in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2019, pp. 1–8.
- [22] R. Pérez-Jove *et al.*, “Network traffic foundation models: A systematic review,” *Comput. Netw.*, vol. 276, Feb. 2026, Art. no. 111998.
- [23] X. Meng, C. Lin, Y. Wang, and Y. Zhang, “NetGPT: Generative pretrained transformer for network traffic,” 2023, *arXiv:2304.09513*.
- [24] X. Lin *et al.*, “ET-BERT: A contextualized datagram representation with pre-training transformers for encrypted traffic classification,” in *Proc. ACM Web Conf. (WWW)*, 2022, pp. 633–642.
- [25] S. Beltiukov *et al.*, “netFound: Principled design for network foundation models,” 2026, *arXiv:2310.17025*.
- [26] D. Wu *et al.*, “NetLLM: Adapting large language models for networking,” in *Proc. ACM SIGCOMM*, 2024, pp. 661–678.
- [27] A. Alkhateeb *et al.*, “DeepSense 6G: A large-scale real-world multimodal sensing and communication dataset,” *IEEE Commun. Mag.*, vol. 61, no. 9, pp. 122–128, Sep. 2023.