Graph Neural Network and Spatial-Temporal Foundation Model Integration for Intelligent Mobile Network Traffic Prediction

Haolan Jiang
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
haoranjiang75@buffalo.edu

Abstract

The rapid growth of mobile data traffic and the evolution toward ultra-dense and heterogeneous network architectures demand a fundamental shift in how network intelligence is realized. This paper presents a system-level investigation of the integration of graph neural networks (GNNs) with spatial-temporal foundation models for intelligent mobile network traffic prediction. Moving beyond algorithmic novelty, the analysis focuses on structural trade-offs, architectural design patterns, data infrastructure requirements, and governance mechanisms that shape the feasibility and trustworthiness of such integrated systems. We examine how GNNs can encode dynamic topological relationships among base stations and user mobility patterns, while large-scale spatial-temporal foundation models capture complex temporal dependencies across multiple scales and network regions. The fusion of these two paradigms opens new pathways for highly accurate and generalizable traffic forecasting, yet it introduces profound challenges related to computational sustainability, model interpretability, fairness across diverse geographic and demographic segments, and alignment with evolving regulatory frameworks. The paper discusses deployment strategies spanning centralized cloud, multi-access edge computing, and on-device inference, highlighting the tension between latency constraints and energy consumption. It further elaborates on data governance imperatives in the context of privacy-preserving, cross-operator learning and on the necessity of rigorous model auditing to ensure robustness against distributional shifts and adversarial perturbations. By synthesizing insights from network engineering, artificial intelligence, and technology policy, this work advances a holistic perspective that can guide the responsible integration of foundation model capabilities into critical communication infrastructures.

Keywords

Graph neural networks; spatial-temporal foundation models; mobile network traffic prediction; model integration; AI governance; sustainable AI.

1. Introduction

The exponential increase in mobile data consumption, driven by high-definition video streaming, immersive extended reality applications, and massive Internet of Things deployments, is pushing existing cellular network management paradigms to their limits. Network operators must anticipate traffic fluctuations with high spatial and temporal resolution to enable predictive resource allocation, dynamic spectrum sharing, and proactive congestion control. In this context, deep learning has emerged as a pivotal technology, with graph neural networks proving especially adept at modeling the irregular spatial structures inherent in urban and rural deployments where base stations form complex adjacency topologies influenced by terrain, building geometry, and user mobility flows [1, 2, 3].

Simultaneously, sequence models for time series forecasting have undergone a dramatic evolution, from recurrent architectures to transformers and, most recently, to large-scale spatial-temporal foundation models pre-trained on vast and diverse datasets [4, 5, 6]. These foundation models promise unprecedented generalization capacity and few-shot adaptability, attributes that are highly attractive for mobile network environments where traffic patterns vary widely across cells, hours, and events. However, the direct application of such models to the networking domain remains fragmented, with most existing efforts either optimizing spatial representations through GNNs or leveraging temporal models without exploiting the structural priors offered by graph-based encodings. A deep system-level integration that fuses the spatial inductive biases of GNNs with the transferable temporal knowledge embedded in foundation models has not yet been thoroughly analyzed from an architectural, governance, and sustainability standpoint.

The mobile networking landscape, shaped by standards such as those produced by the 3rd Generation Partnership Project and the O-RAN Alliance, increasingly envisions AI-native architectures where data-driven control loops operate within strict latency bounds and under stringent energy constraints [16]. The integration of foundation models, which are often fundamentally large and computationally intensive, with graph-based processing pipelines raises novel structural trade-offs. These include questions about the optimal placement of model components across cloud, edge, and far-edge resources, the balance between training cost amortization and inference overhead, and the long-term evolution of network intelligence as traffic dynamics shift with new spectrum allocations and service paradigms. Beyond engineering performance, the deployment of such integrated systems intersects with broader societal concerns related to algorithmic fairness, transparency, and the concentration of technological power in a small number of large model providers. A fragmented approach that optimizes only for prediction accuracy without due consideration for these multidimensional constraints risks creating brittle, resource-hungry, and potentially discriminatory network control systems. This paper therefore adopts a holistic, systems-oriented perspective to investigate how GNNs and spatial-temporal foundation models can be brought together in a manner that is not only technically effective but also governable, sustainable, and fair. The subsequent sections unpack the background, architectural patterns, data infrastructure needs, fairness and robustness imperatives, sustainability considerations, and policy implications of this integration, drawing on a wide range of interdisciplinary literature.

2. Background and Related Work

Early approaches to mobile network traffic prediction often relied on classical time series models or simple feedforward neural networks that treated each cell as an independent data source. The introduction of convolutional and recurrent neural networks enabled joint spatial-temporal modeling, but their regular grid assumptions failed to capture the true irregular topology of cellular deployments. The breakthrough came with the formulation of graph-based learning methods that represent cell towers as nodes and their functional interactions as edges, allowing message passing to propagate information along communication and mobility corridors. The spatio-temporal graph convolutional network opened a new paradigm by integrating graph convolutions with gated temporal convolutions [1], and the diffusion convolutional recurrent neural network further enhanced this by modeling bidirectional diffusion processes on directed graphs [2]. Graph WaveNet subsequently demonstrated that adaptive dependency matrices and dilated convolutions could effectively learn hidden spatial dependencies without a predefined adjacency matrix [3]. These architectures have been widely adopted and extended in the mobile networking domain, where traffic prediction tasks must contend with highly dynamic user populations and event-driven surges [7, 8].

In parallel, the field of time series forecasting has been revolutionized by attention mechanisms and transformer-based architectures capable of capturing long-range temporal dependencies. The Informer model introduced probabilistic attention and self-attention distillation to reduce computational complexity for long sequences [4], paving the way for more scalable temporal models. The concept of a foundation model, originally articulated in the context of natural language processing and later generalized to multiple modalities, has since been transplanted to the time series domain [10]. Lag-Llama exemplifies the ambition to build a single, pre-trained model that can be fine-tuned for a wide range of forecasting tasks across different domains and sampling frequencies [5]. Similarly, TSMixer advanced an all-MLP alternative that demonstrated competitive performance with substantially lower parameter counts [6], highlighting the tension between model scale and practical deployment. A more recent line of work has sought to unify spatial-temporal forecasting under a single framework, with models like UniTS offering universal time series representations that can handle imputation, classification, and forecasting without architectural modification [11]. On the graph side, GraphGPT explored how large language model capabilities can be transferred to graph tasks via instruction tuning, suggesting that graph-based foundation models may not need to be trained from scratch but can inherit semantic reasoning from language-oriented pre-training [12]. The integration of large model paradigms into spatial-temporal traffic intelligence was further exemplified by approaches that incorporate deep language model priors into structured prediction pipelines [13].

Despite these convergent developments, the integration of such large-scale temporal foundation models with graph-based spatial encoders for mobile network traffic prediction remains underexplored at the system level. Existing surveys confirm the efficacy of GNN-based methods for traffic forecasting but largely predate the foundation model era [8]. Research on federated learning for mobile traffic prediction has demonstrated that privacy-preserving collaboration across base stations or operators is feasible, yet these efforts have not been extended to the context of fine-tuning large pre-trained models [9]. The critical gap, therefore, resides not in the absence of usable components but in the absence of a coherent systems methodology that reconciles the scale, latency, energy, and governance requirements of production mobile networks with the capabilities of modern AI building blocks.

3. System Architecture for GNN-Foundation Model Integration

Designing a system architecture that fuses GNNs with spatial-temporal foundation models for mobile traffic prediction requires a careful orchestration of computational modules, data flows, and deployment topologies. At the highest level, the architecture can be conceptualized as a pipeline in which raw network telemetry is ingested and transformed into a dynamic graph representation. Each base station or radio unit corresponds to a node, with features encoding traffic volumes, user counts, active bearers, and context metadata such as time of day and event schedules. Edges capture both physical proximity and functional coupling measured through handover statistics, co-interference patterns, or mobility trajectories. A GNN encoder then processes this graph to produce spatially refined node embeddings that aggregate neighborhood information. These embeddings are subsequently combined with historical temporal sequences and fed into the foundation model component, which functions as a temporal reasoning engine. The foundation model, pre-trained on large corpora of time series data perhaps spanning multiple network domains and even non-telecom sources, extracts multi-scale temporal patterns and projects them into a latent space that can be aligned with the spatial representations learned by the GNN. A fusion module, often a cross-attention layer or a simple concatenation followed by dense layers, merges spatial and temporal signals before generating a probabilistic forecast for multiple future time steps.

The structural trade-offs inherent in this architecture are substantial and must be evaluated against operational constraints. One central tension concerns the decision to keep the foundation model frozen, fine-tune it end-to-end together with the GNN, or employ parameter-efficient adaptation techniques such as adapter layers. A frozen foundation model preserves its generalizable knowledge and reduces the risk of catastrophic forgetting, but it may fail to adapt to domain-specific traffic characteristics such as bursty millimeter-wave propagation effects or the regularity of commuter patterns. Full fine-tuning, by contrast, can yield higher accuracy but demands significant communication and computation resources, pushing the system toward centralized data aggregation, which may conflict with privacy regulations and operator-specific data sovereignty requirements. Intermediate solutions, such as leveraging the foundation model as a feature extractor while training only a lightweight GNN frontend and a prediction head, offer a pragmatic compromise. System designers must also decide whether the GNN and foundation model are trained jointly in a single optimization loop or in a staged fashion, with each component pre-trained on auxiliary objectives. Joint training can enhance alignment between spatial and temporal representations but introduces complex gradient flow dependencies that complicate distributed execution. In geographically distributed mobile networks, data often originates at far-edge sites under strict latency budgets, making it necessary to consider split inference strategies where the GNN processes local graph neighborhoods on-site while the foundation model resides in a regional edge cloud or central data center. Such hierarchical placement can reduce upstream data transfer volumes and improve responsiveness, but it introduces new challenges related to model synchronization and the handling of outdated spatial embeddings when network topology changes.

Scalability and resilience form another critical architectural dimension. Mobile networks are inherently dynamic: cells can be added, decommissioned, or reconfigured, and user behavior can shift abruptly due to large-scale events, disasters, or new service launches. A system integrating GNNs and foundation models must gracefully handle topological evolution without full retraining. This can be approached through continual learning mechanisms that allow the GNN to expand its receptive field as new nodes appear, while the foundation model can be adapted through low-rank updates on incoming data streams. The architecture should also incorporate uncertainty quantification modules that can signal when predictions are unreliable due to distributional shifts, triggering fallback mechanisms based on simpler models or rule-based heuristics. Integration with the broader network intelligence fabric, such as the non-real-time and near-real-time RAN intelligent controllers defined by O-RAN, demands that the model exposes well-defined interfaces for training pipelines, inference serving, and performance monitoring, conforming to the specifications being developed by industry bodies [16]. This interoperability is essential to avoid vendor lock-in and to enable multi-vendor AI ecosystems where operators can mix and match GNN components from one supplier with foundation models from another.

4. Data Infrastructure and Governance

The predictive power of an integrated GNN-foundation model system rests heavily on the quality, coverage, and governance of the underlying data infrastructure. Mobile network traffic data is generated continuously by millions of user equipments across thousands of cells, resulting in streaming telemetry that must be collected, aggregated, and aligned with graph topology metadata. A robust data pipeline must handle variable reporting intervals, missing values due to measurement gaps or equipment failures, and the need for real-time preprocessing to support inference windows that may be as short as a few hundred milliseconds. At the same time, the richness of per-user or per-session data, which would enable fine-grained spatial-temporal modeling, is often inaccessible due to privacy regulations such as the General Data Protection Regulation and similar frameworks in other jurisdictions. Data minimization principles require that only strictly necessary features are collected and that personally identifiable information is stripped before any model training or inference. The tension between data utility for training expressive models and the imperative of user privacy is a fundamental structural constraint that must be designed into the data architecture from the start. Approaches such as federated learning offer one pathway to partially reconcile these competing demands by allowing local model updates to be computed at the network edge and aggregated without raw data leaving the base station premises [9]. However, federated fine-tuning of large foundation models introduces significant upstream communication overhead and may require novel compression and gradient sparsification techniques.

Data governance frameworks must also address provenance, versioning, and consent management. The spatial graph that defines the relational structure among cells is itself a dynamic artifact that evolves with network planning, infrastructure sharing agreements, and temporary deployments such as mobile cells for large events. Changes in graph topology must be versioned alongside the corresponding traffic data to enable reproducible training and auditing. Furthermore, as foundation models are pre-trained on massive external datasets that may include time series from domains as diverse as energy, transportation, and finance, a thorough assessment of data lineage becomes essential. The incorporation of data from unrelated domains can introduce subtle biases or lead to the leakage of non-telecom patterns into network traffic forecasts, with potential consequences for decision-making fairness. Operators must establish clear data-use policies and maintain the ability to trace any given prediction back to the training data sources and pre-training corpora that influenced it.

Cross-operator collaboration, which could substantially improve model generalization by exposing foundation models to a wider array of deployment environments, introduces additional governance complexities. Commercial sensitivities and competitive dynamics often inhibit data sharing, even in anonymized forms. Governance agreements modeled on data spaces or trusted research environments may provide a middle ground, allowing models to be trained on pooled data under strictly audited conditions without direct data exchange. In all cases, the data infrastructure must be instrumented to continuously monitor for distribution shifts, feature drift, and data quality degradation, which can silently erode prediction accuracy and trigger unfair resource allocation decisions across different regions or user groups.

5. Fairness, Robustness, and Model Governance

Integrating GNNs with large foundation models in a critical infrastructure system such as a mobile network raises pressing questions about fairness and robustness that extend well beyond average accuracy metrics. Traffic prediction systems directly inform resource allocation algorithms that determine bandwidth distribution, handover thresholds, and energy-saving mode selections. If the underlying model systematically underestimates traffic in certain geographic areas or for particular demographic populations, the resulting underprovisioning can exacerbate the digital divide by creating zones of degraded service quality. Fairness in this context can be analyzed along multiple axes, including spatial fairness across urban versus rural cells, temporal fairness across peak and off-peak periods, and group-level fairness with respect to user mobility patterns that correlate with socioeconomic status. The pre-training data for foundation models, often sourced from publicly available time series repositories, may underrepresent network traffic patterns from developing regions or from non-mainstream communication behaviors, perpetuating representational biases that cannot be easily corrected during fine-tuning [17]. A deliberate fairness-aware design methodology is therefore required, involving the selection of fairness metrics that align with operator policy objectives, the application of reweighting or adversarial debiasing techniques during training, and the implementation of post-hoc fairness audits as part of a continuous monitoring framework.

Robustness is equally critical. Mobile network traffic is susceptible to sudden perturbations caused by flash mobs, emergency situations, infrastructure failures, and adversarial attacks that inject subtle perturbations into the input data stream with the aim of degrading prediction quality or triggering resource misallocations [18]. GNNs are known to be vulnerable to structural attacks that insert or delete edges in the graph, which could correspond to an attacker compromising the adjacency data or spoofing handover reports. Foundation models, especially those built on transformer backbones, can be sensitive to adversarial noise patterns that are imperceptible in the time domain but cause large deviations in the forecast. The integration of these two components expands the attack surface, as an adversary could target the spatial encoding, the temporal encoding, or the fusion layer. Defense strategies must therefore be multi-layered, incorporating input validation, uncertainty-based anomaly detection, and adversarial training regimes that specifically seek to harden the interface between GNN and foundation model layers. The concept of model governance extends beyond technical robustness to encompass the entire lifecycle of development, deployment, and decommissioning. Clear accountability structures must be established to determine who is responsible when an AI-driven decision leads to a network outage or a discriminatory service pattern. This calls for model cards, regular third-party audits, and alignment with emerging regulatory instruments such as the European Union Artificial Intelligence Act, which classifies AI applications in critical infrastructures as high-risk and subjects them to conformity assessments and transparency obligations [19].

Explainability is a key enabler of both fairness and robustness, allowing network operation teams to interrogate model decisions and identify potential failure modes. While foundation models are frequently criticized for their opacity, explainability techniques such as post-hoc attention visualization, perturbation-based saliency mapping, and concept-based explanations can be adapted to the integrated GNN-foundation setting to reveal whether a prediction relied on legitimate spatial neighborhood effects or on spurious correlations with protected attributes [20]. Such explanations are not merely a research curiosity; they are increasingly becoming a regulatory requirement and a practical necessity for engineers who must trust and act upon model recommendations in high-stakes operational contexts.

6. Sustainability and Deployment Considerations

The sustainability implications of integrating large foundation models into the fabric of mobile network operations deserve careful scrutiny. The training of state-of-the-art foundation models consumes substantial amounts of electrical energy and associated carbon emissions, raising concerns about the net environmental impact of deploying such models at scale across hundreds of operator networks worldwide [15]. While the amortized cost per prediction may be low once a model is trained and shared, the aggregate energy footprint of inference, particularly if models are replicated in every regional data center or if frequent fine-tuning is required to keep them synchronized with local traffic dynamics, can become considerable. This challenge is compounded in mobile networks where base stations and edge servers often operate under power budgets that are tightly constrained by both economic and environmental goals. The architectural choices discussed earlier—frozen versus fine-tuned, centralized versus distributed—directly shape the energy profile of the system. Deploying a full foundation model at a resource-constrained edge node is often infeasible without aggressive compression through quantization, pruning, or knowledge distillation into smaller surrogate models that retain most of the predictive power while drastically cutting floating-point operations per inference.

The system design must therefore navigate a multi-objective optimization landscape where accuracy, latency, energy, and cost are jointly considered. One promising direction is the development of hierarchical model families in which a large, highly accurate foundation model serves as a teacher for a fleet of lightweight student models deployed at the far edge. These students can be specialized per region or per cell cluster and can be periodically updated using feedback from the central teacher in a manner analogous to federated distillation, reducing the need for raw data transfer. The reuse of a single foundation model across multiple tasks—traffic prediction, anomaly detection, energy-saving configuration—further improves the sustainability calculus by replacing the need for many bespoke models with a shared, multi-purpose backbone. Standardization efforts that define open interfaces for model sharing and lifecycle management will be crucial to realizing this vision across a competitive operator landscape, preventing wasteful duplication of computationally intensive pre-training efforts.

Deployment in live network environments also introduces operational sustainability considerations that extend beyond energy. The reliability of model predictions under hardware variability, software updates, and evolving network configurations must be validated through robust continuous integration and continuous delivery pipelines that incorporate automatic rollback mechanisms. The longevity of a deployed model depends on its capacity to adapt to concept drift without requiring frequent, carbon-intensive retraining from scratch. Here, the foundation model paradigm's inherent adaptability can be a double-edged sword: it facilitates rapid fine-tuning but also creates the temptation to retrain more often than strictly necessary. A disciplined approach to model versioning, combined with transparent reporting of the carbon cost associated with each update, can help operators make informed trade-offs between performance gains and environmental impact, aligning with broader corporate sustainability commitments and external reporting requirements.

7. Socio-Technical and Policy Implications

The deep embedding of GNN-foundation model systems into mobile network operations carries significant socio-technical and policy implications that call for anticipatory governance. One structural concern arises from the concentration of expertise and computational resources required to pre-train foundation models at scale, which is currently feasible only for a small number of large technology firms and well-funded research institutions. If the mobile industry comes to rely on a handful of proprietary foundation models as the backbone of its AI strategy, this could create new dependencies and reduce the negotiating power of network operators, potentially stifling innovation and raising barriers to entry for smaller players and operators in developing economies. Countervailing forces include the emergence of open-source foundation model initiatives and the mandate within organizations such as the International Telecommunication Union and O-RAN to develop open, interoperable AI architectures. The tension between proprietary model performance and the systemic benefits of a diverse, competitive ecosystem must be actively managed through procurement practices, technology-neutral regulation, and public investment in open AI research tailored to telecommunications.

Policy frameworks must also address the international dimension of mobile network traffic prediction, as data flows and model updates cross national boundaries. Harmonization of data protection standards, model certification requirements, and incident reporting obligations across jurisdictions is needed to prevent regulatory fragmentation from undermining the global scaling of intelligent network management. The classification of AI-enabled network functions as critical infrastructure components under national security frameworks adds another layer of complexity, as it may restrict the use of models trained or hosted in certain geopolitical contexts. Workforce implications constitute a further socio-technical dimension. The shift toward AI-native network operation demands a reconfiguration of engineering skill sets, blending domain expertise in radio access and core network technologies with competencies in machine learning operations, model debugging, and AI ethics. Proactive investments in training and organizational change management are essential to ensure that the human operators who remain accountable for network performance can effectively oversee and intervene upon increasingly autonomous AI systems. Ultimately, the responsible integration of GNNs and spatial-temporal foundation models into mobile networks is not solely a technical project but a societal one that requires multi-stakeholder dialogue encompassing regulators, operators, vendors, civil society, and the research community.

8. Conclusion

The convergence of graph neural networks and spatial-temporal foundation models represents a transformative opportunity to elevate the intelligence, efficiency, and adaptability of mobile network traffic prediction. This paper has systematically examined the integration challenge through a systems lens, moving beyond algorithmic descriptions to engage with the architectural trade-offs, data governance imperatives, fairness requirements, robustness concerns, sustainability constraints, and policy implications that jointly determine the viability and desirability of such systems. The integration paradigm demands a careful orchestration of spatial encoding layers and temporal reasoning engines deployed across complex, heterogeneous infrastructures. It necessitates robust data pipelines that reconcile privacy and utility, fairness-aware training regimes that prevent service quality disparities, and continuous governance mechanisms that ensure accountability and transparency. Sustainability considerations must be placed at the core of design decisions, incentivizing model compression, hierarchical deployment, and resource sharing. From a policy perspective, the integration path must be shaped by deliberate efforts to foster open ecosystems, align with emerging regulations, and invest in the human capital required to operate AI-native networks. As mobile networks evolve toward sixth-generation architectures in an increasingly connected world, the frameworks established today for integrating deep learning components will determine not only the technical performance but also the trustworthiness and equity of the resulting communication services. The cross-disciplinary perspective advocated here aims to serve as a foundation for future research and standardization activities that place long-term societal value on par with predictive accuracy.

References

1. Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (pp. 3634–3640). AAAI Press.

2. Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In International Conference on Learning Representations.

3. Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. (2019). Graph WaveNet for deep spatial-temporal graph modeling. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (pp. 1907–1913). AAAI Press.

4. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 12, pp. 11106–11115).

5. Garza, A., & Mergenthaler-Canseco, M. (2023). Lag-Llama: Towards foundation models for time series forecasting. arXiv preprint arXiv:2310.08278.

6. Das, A., Kong, S., Sen, R., & Zhou, Y. (2023). TSMixer: An all-MLP architecture for time series forecasting. Transactions on Machine Learning Research.

7. Zhang, J., Zheng, Y., & Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

8. Wang, X., Chen, C., & Min, G. (2021). Spatio-temporal graph neural networks for mobile network traffic prediction: A survey. IEEE Communications Surveys & Tutorials, 23(3), 1808–1835.

9. Xu, J., Zhang, Y., & Wang, Q. (2022). Federated learning for mobile network traffic prediction: A communication-efficient approach. IEEE Transactions on Network and Service Management, 19(2), 1582–1593.

10. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

11. Woo, G., Liu, C., Sahoo, D., Kumar, A., & Hoi, S. (2024). Unified training of universal time series forecasting transformers. In Proceedings of the 27th International Conference on Artificial Intelligence and Statistics.

12. Tang, J., Yang, Y., Wei, W., Shi, L., Su, L., ChengTang, J., Yang, Y., Wei, W., Shi, L., Su, L., & Cheng, Y. (2024). GraphGPT: Graph instruction tuning for large language models. arXiv preprint arXiv:2311.06083.

13. Zhang, C., Zhang, H., Qiao, J., Li, Z., & Alouini, M. S. (2025). TIDES: Traffic Intelligence with DeepSeek Enhanced Spatial Temporal Prediction. IEEE Journal on Selected Areas in Communications.

14. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). Association for Computational Linguistics.

15. Polese, M., Bonati, L., D'Oro, S., Basagni, S., & Melodia, T. (2023). Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges. IEEE Communications Surveys & Tutorials, 25(2), 1376–1411.

16. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery.

17. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations.

18. European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (COM(2021) 206 final). European Commission.

19. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). Association for Computing Machinery.