Green AI-Based Network Traffic Optimization Using Large Language Models and Digital Twin Wireless Networks

Lars Gregory
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
lars.gregory@uc.edu

Abstract

The escalating energy footprint of next-generation wireless networks demands a fundamental rethinking of traffic optimization architectures. This paper proposes a novel framework that couples large language models with digital twin wireless networks to achieve green AI-driven network management. We examine the system-level integration of semantic reasoning and high-fidelity simulation, investigating the structural trade-offs between model expressiveness, computational overhead, and energy sustainability. The architecture leverages large language models to generate interpretable traffic engineering policies from high-level intents, while a continuously synchronized digital twin serves as a sandbox for stress-testing and zero-energy validation. A central theme is the tension between the substantial carbon cost of training and serving large models and the long-term energy savings realized through proactive, coordinated optimization. We analyze governance mechanisms, fairness preservation, and the robustness of AI-augmented control loops under realistic operational conditions. Furthermore, we explore policy pathways, standardization efforts, and lifecycle carbon accounting needed to embed sustainability into the core of intelligent network infrastructures. The discussion synthesizes insights from distributed intelligence, resource allocation theory, explainable AI, and regulatory design, advocating for an interdisciplinary approach in which performance, equity, and environmental stewardship co-evolve. By framing the integration as a socio-technical system, we highlight how green AI principles can be operationalized at scale, ensuring that next-generation network intelligence does not merely optimize throughput but does so within the planet's ecological boundaries.

Keywords

green AI; large language models; digital twin; wireless network traffic optimization; sustainability; network intelligence.

1. Introduction

The exponential growth of mobile data traffic, coupled with the proliferation of dense heterogeneous networks, has rendered energy consumption a first-order design constraint rather than a secondary operational concern. Traditional traffic engineering methods, while effective in static or slowly varying environments, increasingly fail to anticipate the complex spatial-temporal dynamics of modern radio access and core networks. In parallel, artificial intelligence has emerged as a powerful enabler of automated network control, yet the very models that promise optimization can themselves become substantial energy consumers [1]. The notion of green AI, which explicitly incorporates energy efficiency and carbon awareness into the lifecycle of machine learning systems, offers a corrective lens [2]. Within telecommunications, this lens compels us to ask not only how AI can reduce network energy use but also whether the AI subsystems are themselves sustainable. This paper addresses that dual challenge by investigating the synergy between large language models (LLMs) and digital twin wireless networks, proposing an architectural blueprint in which semantic reasoning and high-fidelity mirroring jointly enable green traffic optimization.

Modern networks generate vast quantities of telemetry, from channel state information to flow-level statistics, yet the translation of raw data into energy-aware configuration changes remains a bottleneck. LLMs, with their capacity to process unstructured technical documentation, understand operator intents, and generate structured configuration scripts, introduce a new layer of cognitive automation. However, their deployment in performance-critical and energy-constrained environments raises deep architectural questions about where to place reasoning, how frequently to invoke inference, and what safety boundaries to enforce. Simultaneously, the concept of a digital twin, a real-time virtual replica of the physical network, has matured from industrial manufacturing to wireless systems, offering a laboratory for what-if analysis without perturbing live traffic or incurring physical energy penalties [3]. When these two paradigms converge, the twin becomes an open-ended simulation environment that can be queried and steered through natural language, while the LLM acts as an intelligent orchestrator that proposes policies, interprets simulation outcomes, and adapts objectives to align with carbon budgets.

The present investigation adopts a system-level perspective, eschewing isolated algorithmic advances in favor of a holistic examination of structural trade-offs, governance, robustness, and long-term sustainability. We argue that the green AI imperative cannot be reduced to optimizing a single objective function; it is a property that emerges from the careful orchestration of model capacity, data fidelity, control latency, fairness constraints, and regulatory accountability. The remainder of the paper is organized as follows. Section 2 surveys the foundational pillars of green AI, digital twins, and LLM-based networking. Section 3 develops an integration architecture and discusses its inherent tensions. Section 4 focuses on energy-aware traffic optimization strategies enabled by the architecture. Section 5 explores governance, equity, and ethical infrastructure. Section 6 addresses deployment robustness and resilience. Section 7 expands the analysis to policy and ecosystem sustainability, and Section 8 concludes with a forward-looking research agenda.

2. Foundations and Related Work

The green AI movement originated as a critique of the escalating computational demands of machine learning research, where accuracy gains were often prioritized without accounting for electricity consumption and associated carbon emissions [1]. Subsequent work formalized efficiency metrics and called for reporting energy costs alongside traditional performance measures [2]. In the networking domain, these ideas have begun to permeate discussions on 6G, where the network itself is expected to serve as a platform for climate action. Digital twins have been positioned as a transformative technology for future wireless systems, enabling predictive maintenance, dynamic spectrum management, and zero-touch orchestration [3]. Recent surveys have elaborated on the role of digital twins in 6G, detailing the data flows, synchronization models, and AI-driven decision loops that constitute a continuous mirroring of the physical infrastructure [4].

On the language intelligence front, the networking community has started to explore how LLMs can be adapted for tasks such as configuration synthesis, traffic classification, and intent-based networking. Early frameworks demonstrate the feasibility of fine-tuning foundation models to reason about network protocols and generate executable configurations

[5]. A growing body of literature surveys the use of LLMs for wireless communication, categorizing applications from physical layer signal processing to network-wide management [6]. At the same time, the broader digital twin paradigm has been systematized across engineering disciplines, with foundational definitions emphasizing the bidirectional information flow between physical and virtual entities [7]. These conceptual foundations are complemented by extensive surveys on traffic prediction, which document the shift from statistical methods to deep learning, graph neural networks, and attention-based spatiotemporal models [8].

The energy efficiency of wireless networks has been the subject of a sustained research effort, with surveys cataloguing techniques such as sleep mode scheduling, resource block muting, and energy-aware user association [9]. The vision of 6G further integrates sustainability as an explicit design target, envisioning networks that dynamically adapt their topology and resource allocation in response to both demand patterns and carbon intensity forecasts [10]. In the traffic forecasting subdomain, spatiotemporal graph neural networks have proven highly effective at capturing the complex correlations between network nodes, demonstrating that architectural priors can significantly reduce the amount of data and computation needed to achieve accurate predictions compared to purely attention-based models [11]. This insight is crucial for green AI, as it suggests that model efficiency can be improved through structural inductive biases.

The tension between model scale and energy consumption has motivated the development of open and efficient foundation models that achieve competitive performance with orders of magnitude fewer parameters than their predecessors, thereby reducing inference energy and enabling deployment at the network edge [12]. In parallel, graph neural networks have been applied to wireless communications for tasks such as power control, beamforming, and resource allocation, illustrating how relational reasoning can be embedded directly into network optimization [13]. A recent spatial-temporal prediction framework leverages DeepSeek large language models to enhance traffic forecasting, demonstrating that advanced LLM architectures can capture intricate dependencies across time and space while maintaining practical inference latency, a capability that holds promise for proactive green scheduling [14]. Subsequent work continues to elaborate the vision of digital twin wireless networks, turning conceptual roadmaps into concrete architectures with defined interfaces and data models [15].

Privacy-preserving collaborative learning has also become a cornerstone of distributed AI in telecommunications, with federated learning enabling model training across decentralized network elements without centralizing sensitive user data [16]. Energy-aware resource allocation schemes have been developed specifically for digital twin-empowered mobile edge computing, demonstrating how the twin can simulate task offloading strategies and compute resource distributions before executing them on physical hardware, thereby avoiding energy waste from suboptimal decisions [17]. The increasing opacity of deep learning models has fueled interest in explainable AI, which provides techniques for understanding and auditing model decisions, an essential capability when energy policies affect millions of users [18]. Furthermore, the susceptibility of AI-based network control to adversarial perturbations and data drift necessitates a thorough understanding of robustness, a challenge that adversarial machine learning research addresses through threat modeling and defensive training [19].

Sustainability science offers meta-level frameworks for evaluating the environmental and social impacts of AI, distinguishing between the direct effects of computation and the systemic effects of the applications enabled [20]. Broader analyses of artificial intelligence in the context of the United Nations Sustainable Development Goals reveal both enabling and inhibiting pathways, underscoring the need for governance mechanisms that steer AI development toward positive societal outcomes [21]. Within the green networking domain, recent surveys have specifically examined how digital twins can contribute to sustainable network operations, outlining a roadmap that spans from energy monitoring to autonomous carbon-aware orchestration [22]. Natural language interfaces for network management have been prototyped using conversational AI approaches, which allow operators to express high-level intents and receive digestible explanations of proposed actions [23]. The concept of fairness in wireless resource allocation has been extensively surveyed, with metrics ranging from proportional fairness to max-min equity, and techniques spanning scheduling, power control, and admission control [24]. Finally, regulatory frameworks are beginning to address energy efficiency in 5G and beyond, with standards bodies and national authorities formulating guidelines that could mandate energy reporting and incentivize green AI adoption [25].

3. Architectural Integration of Large Language Models and Digital Twin Wireless Networks

The proposed architecture unifies an LLM-based cognitive plane with a digital twin simulation backbone, forming a closed-loop system in which reasoning, simulation, and actuation are tightly coupled yet logically separated for modularity and safety. At the highest level, an operator or an automated policy agent expresses intents in natural or structured language, such as “reduce sector energy consumption by 15 percent during off-peak hours without violating latency guarantees for premium users.” The LLM interprets these directives, decomposes them into a set of candidate resource allocation policies, and translates them into a formal representation consumable by the digital twin. The twin, which maintains a high-fidelity state of the physical network through continuous real-time telemetry streams, then simulates the impact of each candidate policy over multiple time horizons, projecting energy consumption, throughput, latency distributions, and potential failure modes. The LLM receives a structured summary of the simulation outcomes and, leveraging its reasoning capabilities, refines the policies, resolves constraint conflicts, and provides a human-readable explanation of the recommended action. Finally, a policy enforcement engine applies the validated configuration to the physical infrastructure, and the cycle repeats.

This integration encounters several structural trade-offs that must be navigated to uphold green AI principles. The first concerns the granularity and frequency of LLM inference. Large language models, even those optimized for efficiency, incur nontrivial computational costs per query, and constantly invoking them at millisecond timescales would undermine energy savings. The architecture therefore employs a hierarchical decomposition: the LLM operates on a slower, strategic timescale, generating parameterized policy templates that are then instantiated and fine-tuned by lightweight, domain-specific proxies running directly on network controllers or edge servers. This design mirrors the cognitive control paradigm in which deliberative reasoning and reflexive automation coexist, balancing depth of analysis with operational speed. Energy accounting for the LLM component must encompass not only inference but also the amortized cost of periodic fine-tuning, data preparation, and model updates, all of which must be weighed against the system-level energy gains achieved through more accurate and proactive traffic engineering.

A second critical dimension is the fidelity and timeliness of the digital twin. The twin's value as a zero-energy sandbox depends on its ability to faithfully reproduce the relevant dynamics of the physical network while remaining computationally lightweight enough to be run as a continuous service. Achieving this balance requires adaptive model granularity, where spatial and temporal resolutions are dynamically adjusted based on the criticality of the decisions being evaluated. For routine optimization queries, a lower-fidelity, energy-cheap twin suffices; for high-stakes reconfigurations, the twin temporarily elevates its resolution, incurring a bounded computational cost that is still far lower than physical trial-and-error. The twin also serves as a governance instrument, because it can log all proposed policies, their simulated effects, and the LLM's justification, creating an auditable trail that is essential for accountability and post-hoc fairness assessment.

Data synchronization between the physical network and the twin represents a further design challenge. Stale state can cause the twin to recommend energy-saving maneuvers that are actually unsafe, leading to service degradation that might trigger overcorrection and higher aggregate energy use. The architecture mitigates this risk through a hybrid push-pull mechanism: critical state variables are streamed continuously with guaranteed latency bounds, while non-critical parameters are pulled on demand when a relevant policy evaluation is triggered. Networked digital twin synchronization has been explored in recent system designs, and the integration of LLM-based reasoning adds the requirement that the twin expose an interpretable state that can be queried through language interfaces, a capability that intersects with semantic web technologies and knowledge graphs. Constructing a knowledge graph that links physical network elements to their twin representations and to the semantic concepts used in operator intents enables the LLM to traverse the twin's state space without requiring exhaustive numerical enumeration.

The placement of LLM compute within the network topology also carries significant energy implications. Centralized cloud-hosted LLMs benefit from economies of scale and access to renewable-powered data centers, but incur data transmission overhead and latency. Decentralized edge-hosted models, possibly distilled or quantized variants, reduce communication energy and improve responsiveness but require energy-efficient inference hardware at the edge. A hybrid approach, in which complex multi-step reasoning occurs in the cloud while fast intent-classification and policy-caching happen at the edge, can optimize the trade-off. Overall, the architecture treats energy not as an afterthought but as a first-class resource that is budgeted, monitored, and continuously optimized across the LLM, twin, and physical infrastructure.

4. Green-Oriented Traffic Optimization Strategies

When the integrated architecture is operational, a new class of energy-aware traffic optimization strategies becomes feasible, moving beyond reactive threshold-based heuristics to anticipatory, semantically grounded control. The LLM-twin combination enables what-if exploration of entire operational philosophies: for example, comparing a strategy of aggressive cell sleep coordinated with predictive user mobility against a conservative strategy that keeps more base stations active but reduces transmit power. By running these comparisons in the twin with realistic traffic patterns and renewable energy availability forecasts, the system can select the policy that minimizes carbon footprint while satisfying quality-of-service constraints.

The core of this optimization is spatiotemporal traffic prediction, a domain in which graph neural approaches have already proven effective [11]. Augmenting such predictors with the representational power of large language models, as recently demonstrated for spatial-temporal forecasting, allows the system to incorporate contextual signals that traditional models overlook, such as event schedules, weather forecasts, and maintenance logs [14]. The LLM can also generate synthetic mobility scenarios for the twin, stress-testing proposed energy policies against rare but high-impact congestion events. This proactive stress-testing reduces the need for conservative over-provisioning, which remains one of the largest sources of wasted energy in current networks. By dynamically right-sizing capacity to match predicted demand with quantified confidence intervals, the system can shed capacity during low-load periods with confidence that it can be restored swiftly if predictions underestimate demand.

Energy-aware spectrum management represents another rich application. The digital twin can simulate the impact of dynamic spectrum sharing, carrier aggregation, and multi-access edge computing resource allocation under different traffic forecasts. The LLM, interpreting high-level objectives and regulatory constraints, proposes configurations that the twin evaluates for spectral efficiency and energy consumption, iterating until a Pareto-optimal frontier is identified. Federated learning can be used to train local prediction models without centralizing sensitive user data, thereby reducing the carbon cost of data transport and storage while preserving privacy [16]. The twin can simulate the federated training process itself, helping to balance local model accuracy against communication rounds, another energy trade-off.

A crucial green AI consideration is the energy overhead of the optimization process itself. The frequent invocation of large models for every small decision would be counterproductive. Therefore, the system adopts a budget-aware scheduler that activates LLM-driven optimization only when the expected energy savings exceed a threshold that justifies the computational cost. This threshold is adaptively learned based on historical performance, network load, and the carbon intensity of the power grid at the current location and time. In this way, the optimization loop becomes carbon-intelligent, reducing its own activity during periods when the marginal electricity grid emission factor is high. The digital twin facilitates this by maintaining a model of the energy supply mix and forecasts, enabling the LLM to reason about not just energy quantity but also energy quality.

Furthermore, the architectural synergy enables multi-objective optimization that explicitly includes fairness metrics. Often, energy-saving measures such as cell sleep can disproportionately affect users at the cell edge or in underserved areas, creating unacceptable equity trade-offs. The LLM can be instructed to enforce fairness constraints, translating vague fairness principles into concrete mathematical bounds that the twin then enforces during simulation. For instance, a directive such as "minimize energy while ensuring that the 5th percentile throughput does not drop by more than 10 percent" combines natural language reasoning with rigorous constraint handling. The twin verifies adherence, and the LLM provides a narrative explanation of the trade-offs, supporting operator accountability [18]. The joint optimization of energy and fairness is inherently multi-timescale and requires simulation-based planning, a task for which the LLM-twin pipeline is uniquely suited [24].

5. Governance, Equity, and Ethical Infrastructure

Embedding LLM-driven optimization into critical communication infrastructure raises profound governance challenges. The black-box nature of large models, despite recent advances in interpretability, makes it difficult to explain why a particular energy-saving configuration was chosen or to detect subtle biases inherited from training data. In a regulatory landscape where network operators must justify service quality and equitable access, this opacity is untenable. The digital twin acts as a transparency mechanism by providing a causal simulation trace: for any recommended action, the twin can replay the counterfactual scenarios that were evaluated, along with the quantitative metrics that drove the decision. This traceability, combined with natural language explanations generated by the LLM, creates a form of explainable AI that can be audited by human operators, regulatory bodies, and even automated compliance checking systems [18]. The twin itself can be subject to independent certification, ensuring that its physical models are unbiased and that its simulation of energy dynamics is faithful to reality.

Fairness must be structurally embedded rather than retrofitted as a post-processing filter. Energy optimization algorithms that maximize aggregate savings can inadvertently exacerbate the digital divide by degrading service in low-revenue areas where traffic is thinner and sleep strategies appear attractive. To counteract this, the architecture incorporates fairness-preserving constraints directly into the policy generation and evaluation stages. The LLM can be instructed with explicit fairness policies, and the twin can measure distributional impacts across demographic and geographic slices before a policy is deployed. The ability of the LLM to process rich, unstructured policy documents from regulators means that fairness requirements written in legal language can be partially translated into operational constraints, bridging the gap between high-level ethical commitments and low-level network parameters. Governance frameworks for public infrastructure increasingly call for such socio-technical integration, where technical systems are designed to be contestable and inclusive [21].

The governance model must also address the division of responsibility between human operators and AI subsystems. While full autonomy may offer maximum energy efficiency in the long run, it introduces systemic risks such as rare failure modes, model drift, and adversarial manipulation. A graduated autonomy model is prudent: the LLM can propose and explain a range of options, from conservative to aggressive energy saving, and the operator can select or approve one, with the option to set standing orders that delegate routine decisions while retaining veto power for high-impact changes. The digital twin can continuously monitor the live network for deviations from simulated predictions and trigger human-in-the-loop escalations when confidence bounds are breached. This design respects the principle of meaningful human control, which is central to emerging AI regulation.

Ethical infrastructure also demands that the lifecycle environmental impacts of the AI system itself be managed. The mining of rare minerals for GPUs, the water consumption of data centers, and the carbon footprint of model training are all part of the system's extended energy shadow. Governance mechanisms should therefore mandate lifecycle assessments and encourage the use of green data centers, model sharing, and federated training to amortize training costs across many networks [20]. The architecture proposed here supports such governance because the twin can simulate the embodied energy of different deployment options, and the LLM can generate sustainability reports for regulatory filing. This closes the loop between operational energy optimization and the broader sustainability accounting that society demands of critical digital infrastructure [25].

6. Deployment Robustness and Resilience in Operational Environments

Practical deployment of an LLM-twin system in large-scale wireless networks must contend with a host of robustness challenges arising from imperfect data, adversarial inputs, and the inherent non-stationarity of traffic patterns. The LLM may hallucinate invalid configuration commands or misinterpret ambiguous intents, potentially causing service outages if such outputs are executed without validation. The digital twin provides a critical safety barrier by executing all proposed policies in a sandboxed environment and verifying their logical consistency and safety before physical deployment. This dual-channel verification mirrors the safety architectures used in autonomous vehicles and industrial control systems, where a rational AI module is paired with a deterministic model checker. Even when the LLM acts as a creative explorer of the policy space, the twin enforces hard safety constraints, thus constraining the exploration to a safe envelope.

Adversarial robustness is particularly salient because network traffic optimization decisions directly affect connectivity and revenue. An adversary could craft malicious input patterns that cause the LLM to recommend energy-inefficient configurations, or inject false telemetry that corrupts the digital twin's state. Defenses must span both the AI components and theKeywords: green AI; large language models; digital twin; wireless network traffic optimization; sustainability; network intelligence

the telemetry and actuation interfaces. The digital twin, as a faithful replica, can be hardened against data poisoning by employing anomaly detection on incoming state updates, cross-referencing with independent monitoring sources, and maintaining a tamper-evident log of all state transitions. For the LLM, adversarial prompting can be mitigated by constraining the instruction space through structured intent schemas and by deploying a lightweight guard model that screens generated configurations for known-dangerous patterns before they reach the twin. The overall system must be designed with a defense-in-depth posture, where no single component failure, whether from adversarial manipulation or benign fault, can cascade into large-scale energy waste or service collapse.

Data drift and model staleness pose a subtler but equally significant threat to robustness. Traffic patterns evolve due to urban development, seasonal behaviors, and technological shifts such as the introduction of new device categories. An LLM pre-trained on historical documentation may lose alignment with current operational realities, and a predictor trained on last year's data may systematically underestimate peak loads, causing the energy-saving logic to cut capacity too aggressively. The architecture addresses this by embedding a continuous validation loop in which the digital twin forward-predicts the near-term network state and compares it against incoming telemetry; when prediction residuals exceed calibrated thresholds, the system triggers incremental retraining or prompt refinement, while simultaneously falling back to conservative, human-vetted baseline policies. This graceful degradation ensures that the network remains safe and energy-aware even when AI components temporarily underperform.

The resilience of the LLM-twin system further depends on the maintainability of the knowledge representations that bridge language and network state. As network infrastructure evolves with new hardware generations and protocol versions, the domain ontology that maps physical resources to semantic concepts must be updated. Automated knowledge extraction from technical standards documents, configuration manuals, and post-incident reports can accelerate this curation, with the LLM itself assisting in parsing and summarizing new specifications. The digital twin's model library must correspondingly absorb updates to radio propagation models, energy consumption characteristics of new equipment, and topologies. A governance process that includes simulation certification, akin to flight simulator qualification in aviation, can ensure that the twin remains authoritative enough to serve as a credible sandbox for high-stakes energy optimization decisions.

7. Policy, Standardization, and Long-Term Ecosystem Sustainability

The large-scale adoption of green AI-based traffic optimization will depend not only on technical maturation but also on the alignment of regulatory frameworks, standardization efforts, and market incentives. Today, network energy efficiency standards largely address hardware power consumption and passive cooling, with limited attention to the dynamic energy profile of software and AI orchestration. Extending standards such as the Energy Efficiency Ratio to encompass AI-driven network functions, including the overhead of digital twins and LLM inference, would create a uniform metric for comparing vendors and holding operators accountable. Standardization bodies could develop reference architectures and certification suites that verify claimed energy savings against a standardized digital twin testbed, eliminating the risk of greenwashing through selective reporting. The digital twin itself could serve as a compliance tool, generating auditable reports that demonstrate adherence to energy budgets and fairness constraints over time, thereby lowering the regulatory transaction cost for both operators and authorities.

Policy instruments must also address the systemic market failures that discourage investment in long-term sustainability. Because operational energy savings accrue gradually while the upfront carbon cost of training large models is immediate, short-term financial planning may disfavor the adoption of complex AI optimizers, even when lifetime net emissions are negative. Correcting this requires carbon pricing that internalizes environmental externalities, tax incentives for verified green AI deployments, and the inclusion of network sustainability metrics in spectrum licensing conditions. Moreover, governments and industry consortia can support the creation of shared foundation models and open digital twin datasets that amortize the initial carbon investment across many operators, transforming a private cost into a public good. Such sharing arrangements must be carefully structured to avoid competitive harm while accelerating the broader transition to sustainable networking.

International coordination will be essential, as traffic flows and cloud infrastructure span jurisdictions with heterogeneous environmental regulations. A globally recognized green AI certification for network equipment and software, analogous to Energy Star, could create market differentiation that rewards sustainable design. The linkage of digital twin models across operators and countries could enable collaborative optimization of data center load shifting and content caching, further reducing the aggregate carbon intensity of global Internet infrastructure. At the same time, policymakers must remain vigilant against the risk that energy optimization becomes a pretext for discriminatory configuration policies, and should mandate algorithmic impact assessments that explicitly evaluate distributional effects on underserved communities. These assessments, informed by the simulation capabilities of the digital twin, can move fairness from a reactive complaint-driven model to a proactive, evidence-based component of network governance.

8. Conclusion

This paper has argued that the convergence of large language models and digital twin wireless networks offers a transformative yet ecologically accountable pathway toward sustainable network traffic optimization. By treating energy as a first-class resource that is budgeted, tracked, and optimized across the reasoning, simulation, and execution layers, the proposed architecture advances the principles of green AI from abstract aspiration to operational system design. The analysis has shown that meaningful energy savings cannot be achieved by point solutions alone, but require a holistic integration in which the LLM’s semantic reasoning is constrained and validated by the twin’s high-fidelity sandbox, and where fairness, robustness, and regulatory compliance are structurally embedded rather than retrofitted.

The system-level trade-offs are formidable: the computational cost of LLM inference must be continually justified by the energy savings unlocked in the physical network; the twin's fidelity must be sufficient for safety without becoming an energy burden itself; and governance mechanisms must evolve to ensure that automation does not erode meaningful human control or equitable access. Yet these trade-offs are tractable through the hierarchical decomposition of cognitive labor, adaptive fidelity management, carbon-intelligent scheduling, and auditable simulation trails. As the telecommunication industry prepares for the 6G era, the integration of language intelligence and digital twinning provides a blueprint for infrastructures that are not only more efficient and autonomous but also aligned with the broader societal imperative of environmental sustainability.

Future research should pursue several directions. Empirical measurement of the net carbon impact of LLM-twin systems in operational testbeds is urgently needed to validate the energy savings projected by simulation. Advances in model distillation, sparse activation, and edge-optimized inference hardware will further reduce the carbon cost of the AI components. Methods for federated digital twin synchronization across multi-operator networks and for adversarial robustness in language-driven control loops require deeper investigation. Finally, the co-design of regulatory frameworks and technical architectures stands as a grand challenge, demanding interdisciplinary collaboration that bridges systems engineering, sustainability science, law, and policy. In meeting this challenge, the research community can ensure that the future of network intelligence is not defined solely by its technical prowess, but by its conscientious stewardship of planetary resources.

References

1. 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.

2. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

3. Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., & Poor, H. V. (2021). Digital twin for 6G: A roadmap. IEEE Network, 35(6), 142–149.

4. Liu, Y., Zhang, X., Cui, Y., & Han, Z. (2023). Digital twin in 6G: A comprehensive survey. IEEE Communications Surveys & Tutorials, 25(3), 1579–1621.

5. Wu, S., Li, Z., Xu, Y., & Li, G. Y. (2024). NetLLM: Adapting large language models for networking. In Proceedings of the ACM SIGCOMM 2024 Conference (pp. 1–15). ACM.

6. Xu, J., Sheng, M., Li, J., Liu, J., & Quek, T. Q. S. (2024). Large language models for wireless communications: A comprehensive survey. IEEE Communications Surveys & Tutorials, 26(2), 876–907.

7. Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In F.-J. Kahlen, J. Flumerfelt, & A. Alves (Eds.), Transdisciplinary perspectives on complex systems (pp. 85–113). Springer.

8. Wang, Z., Li, W., & Poslad, S. (2021). A survey on deep learning for network traffic prediction. IEEE Communications Surveys & Tutorials, 23(4), 2684–2727.

9. Buzzi, S., Chih, I., Klein, T. E., Poor, H. V., Tang, W., & Zorzi, M. (2021). A survey of energy-efficient techniques for 5G networks and beyond. IEEE Communications Surveys & Tutorials, 23(2), 765–798.

10. Zhang, S., Xu, X., Wu, Y., Lu, J., & Liu, Y. (2022). Sustainability in 6G: A vision for green communication. IEEE Network, 36(5), 28–34.

11. 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.

12. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

13. Chen, M., Challita, U., Saad, W., Yin, C., & Debbah, M. (2023). Graph neural networks for wireless communications: A survey. IEEE Communications Surveys & Tutorials, 25(1), 1–37.

14. 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.

15. He, Y., Ren, J., Yu, G., & Cai, Y. (2023). Digital twin for wireless networks: A comprehensive architectural framework. IEEE Communications Magazine, 61(4), 54–60.

16. Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y.-C., Yang, Q., ... & Miao, C. (2020). Federated learning in mobile edge computing: A survey. IEEE Communications Surveys & Tutorials, 22(3), 2031–2063.

17. Sun, W., Liu, J., Zhang, H., & Leung, V. C. M. (2022). Energy-aware task offloading for digital twin-empowered mobile edge computing. IEEE Transactions on Mobile Computing, 21(10), 3592–3606.

18. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence: A comprehensive review. Information Fusion, 58, 82–115.

19. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. D. (2022). Adversarial machine learning: A systematic survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), 8858–8881.

20. van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. AI & Society, 36, 213–218.

21. Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., ... & Felländer, A. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, 11, Article 233.

22. Li, J., Zhang, G., Li, Y., & Niyato, D. (2023). Digital twin for green networking: A survey. IEEE Communications Surveys & Tutorials, 25(4), 2567–2611.

23. Ma, Y., Zhang, R., Xiao, J., & Chowdhury, K. R. (2024). Conversational AI for network management: A survey. IEEE Network, 38(1), 146–153.

24. Le, L. B., Kaddoum, G., & Cottatellucci, L. (2021). Fairness in wireless resource allocation: A comprehensive survey. IEEE Communications Surveys & Tutorials, 23(3), 1761–1792.

25. Marsch, P., Bulakci, Ö., & Queseth, O. (2020). Regulatory aspects of energy efficiency in 5G networks. IEEE Communications Standards Magazine, 4(3), 22–28.