Large Language Model-Driven Intent-Aware Resource Allocation for 6G Semantic Communication Networks

Walid Craig
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
hellowalid@oregonstate.edu

Tylan Parker
Department of Computer Science, University of Houston, Houston, TX, USA.
dylan618@uh.edu

Krish Banerjee
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
contactkrish@binghamton.edu

Abstract

The emergence of sixth-generation wireless systems redefines communication through semantic-aware architectures that prioritize meaning and task effectiveness over raw bit transmission. Simultaneously, large language models have demonstrated remarkable capabilities in interpreting ambiguous human intent expressed in natural language. This paper presents a comprehensive system-level investigation into the convergence of these two paradigms, proposing an intent-aware resource allocation framework for 6G semantic communication networks that leverages large language models as the core intent interpretation engine. Departing from conventional optimization formulations, we examine the structural trade-offs involved in embedding powerful generative language understanding within the network control plane. The discussion spans architectural decomposition, cross-layer semantic mapping, dynamic resource orchestration, and governance challenges. We analyze how an intent-driven, language-model-mediated control loop can translate high-level user goals into semantically coherent resource assignments across radio, computation, and content dimensions. The study further investigates tensions between inference latency and resource granularity, between centralized intelligence and edge autonomy, and between model adaptability and operational determinism. Sustainability, fairness, trustworthiness, and regulatory alignment are examined as integral design requirements rather than afterthoughts. By situating large language models at the nexus of semantic intent interpretation and infrastructure orchestration, the paper articulates a holistic research agenda that informs future 6G standardization, network operations, and the responsible integration of generative artificial intelligence into critical communication infrastructures.

Keywords

6G, semantic communication, large language models, intent-based networking, resource allocation, network slicing, generative AI governance, sustainability.

1. Introduction

The trajectory of mobile communication systems has evolved from a predominantly bit-centric design philosophy toward architectures that account for the meaning, context, and purpose of information exchange. Sixth-generation networks are expected to realize this shift through semantic communication, wherein transmitters and receivers jointly optimize the transmission of relevant semantic features rather than raw symbols. At the same time, network management is undergoing a transformation toward intent-based operation, which allows stakeholders to express desired outcomes in declarative terms without specifying low-level configuration parameters. The convergence of these two transformations opens an unprecedented design space in which resource allocation decisions must simultaneously respect semantic fidelity, task effectiveness, and high-level business or user intents. Large language models (LLMs) offer a powerful mediating layer capable of parsing natural-language intents, reasoning about context, and mapping abstract goals into actionable semantic resource policies. This paper provides a detailed, cross-disciplinary analysis of how LLM-driven intent-aware resource allocation can be designed, deployed, and governed within 6G semantic communication networks.

Recent visions for 6G emphasize extreme connectivity, integrated sensing and communication, pervasive artificial intelligence (AI), and sustainability by design [1, 2, 3]. The transition from 5G's enhanced mobile broadband and ultra-reliable low-latency communication toward fully immersive, holographic, and tactile applications compels a fundamental rethinking of the communication stack. Semantic communication has been proposed as a pillar of this rethinking because it compresses and transmits only the information that matters for a given task, thereby achieving dramatic improvements in spectral and energy efficiency [4, 5]. Meanwhile, machine learning techniques are increasingly embedded at all layers, turning the network into a distributed learning fabric [6]. The International Mobile Telecommunications 2030 framework further articulates key usage scenarios that include immersive communication, massive communication, and hyper-reliable and low-latency communication, alongside enabling technologies such as integrated AI [7]. In such an environment, static or purely model-driven resource allocation becomes insufficient because user and application semantics constantly shift. Intent-based networking, originally developed within autonomous systems and software-defined control loops, has been identified as a promising paradigm for 6G to manage this complexity by allowing intent to be expressed in a human-understandable manner and then automatically translated into network actions [8]. The recent availability of transformer-based LLMs with broad world knowledge and few-shot reasoning capabilities creates an opportunity to place natural language understanding directly inside the intent-to-policy translation pipeline. When combined with semantic communication frameworks, LLMs can not only interpret what a user wants but also reason about which semantic representations best serve that intent under current network conditions.

Nevertheless, integrating LLMs into the performance-critical control plane of 6G networks raises profound system-level questions that go far beyond accuracy benchmarks. These include the repercussions of inference latency on real-time resource schedulers, the energy footprint of large models in sustainable network deployments, the explainability and auditability of LLM-driven decisions in regulated sectors, and the governance structures needed to prevent intent manipulation or model drift. This paper addresses such questions from a holistic systems perspective, deliberately avoiding narrow algorithmic formulations. Instead, it unfolds a layered architecture discussion that spans intent capture, semantic decomposition, resource brokering, and adaptive governance. The contributions are conceptual and architectural, providing a high-resolution map of trade-offs, design tensions,

and policy implications that researchers, standardization bodies, and operators must confront as they move toward LLM-enhanced semantic 6G infrastructures.

2. Background and Motivation

The progression toward fully semantic-aware networks requires re-examining the relationship between information source, channel, and receiver. In classical communication theory, the physical layer is agnostic to meaning, while semantic communication introduces a semantic encoder that preserves task-relevant information and a semantic decoder that reconstructs meaning at the receiver, often leveraging shared background knowledge [4]. This model aligns naturally with machine learning pipelines, where deep neural networks jointly learn to compress and reconstruct semantic representations tailored to specific downstream tasks [5]. In parallel, the concept of machine learning in the air has generalized the idea that wireless devices and base stations can collaboratively execute inference and training [6]. These advances push resource allocation beyond the traditional metrics of rate, power, and bandwidth toward optimizing semantic value and task utility. As a result, the resource allocation problem becomes intrinsically coupled with application-layer semantics, necessitating a decision layer that understands both the dynamic physical environment and the evolving intent of users and services.

The intent-based networking paradigm offers such a decision layer. Originating from the desire to reduce operator cognitive load and configuration errors, intent-based systems allow operators to specify "what" is desired rather than "how" to achieve it, and rely on closed-loop automation to maintain the desired state [8, 9]. In a 6G semantic context, intents might range from "ensure high immersion quality for a remote collaboration session" to "minimize energy consumption across all surveillance cameras while preserving event detection accuracy." These intents are expressed in natural or constrained language and must be translated into concrete semantic communication parameters, slice templates, radio resource assignments, and edge-computing resource allocations. Conventional rule-based or simple machine learning classifiers struggle with the open-ended, compositional, and context-dependent nature of such intents. It is here that LLMs, with their capacity to handle nuance, resolve ambiguity, and draw analogies, become systemically interesting.

The rise of large-scale transformer models such as GPT-3, GPT-4, LLaMA, and their instruction-tuned variants has demonstrated that language understanding can be operationalized through prompting and retrieval-augmented generation without task-specific fine-tuning in every instance [10, 11, 12]. These models excel at parsing complex instructions, summarizing long documents, and performing chain-of-thought reasoning, attributes that directly apply to decomposing an abstract intent into a hierarchy of sub-goals and constraints. Their deployment in a network control context, however, raises issues of latency, determinism, and resource consumption that differ markedly from consumer chatbots. Therefore, understanding the motivation requires recognizing both the potential of LLMs to handle the semantic richness of 6G intents and the necessity of adapting their deployment architectures for network-grade reliability.

3. System Architecture of Intent-Aware Semantic Resource Allocation

We propose a reference architecture that situates the LLM within a broader intent management and semantic resource orchestration framework. The architecture consists of five interacting functional domains: intent capture and dialogue, intelligent intent interpretation engine, semantic knowledge base, semantic resource broker, and cross-domain infrastructure

controllers. These domains are logically distinct but may be physically distributed across edge clouds, regional data centers, and core network functions.

The intent capture domain provides a multi-modal interface through which human users, enterprise orchestrators, or automated vertical application managers can express connectivity and computation intents. The interaction model is designed as a conversational loop, allowing clarification, negotiation, and refinement. For example, a factory supervisor might initially express an intent as "maintain real-time control loop for all robotic arms, tolerant to occasional sensor failures," and the system may probe for additional details such as acceptable failure response latency or fallback behavior, using the LLM as a conversational agent. This dialogue structure is essential: in complex environments, intent is rarely fully specified a priori, and iterative elicitation mirrors the reasoning patterns that LLMs support naturally.

The intent interpretation engine, centered on an LLM backbone, transforms the refined natural-language intent into a structured semantic intent graph. This graph encodes service objectives, performance constraints, spatial and temporal scope, priority levels, and inter-dependency relations with other ongoing intents. The engine consults a semantic knowledge base that stores domain-specific ontologies of semantic communication tasks, pre-computed semantic encoder–decoder profiles, network capability models, and historical intent–resource mappings. Taking inspiration from retrieval-augmented generation approaches, the engine dynamically retrieves relevant knowledge to ground the LLM's output, reducing hallucination and improving factuality. This step is critical because the output must be interpretable and verifiable by downstream policy engines that lack the flexibility of natural language.

The semantic resource broker translates the intent graph into a set of competing yet compatible resource candidates. Unlike traditional schedulers that operate on bit-rate queues, the broker works with semantic flows, each characterized by a semantic distortion–rate–utility surface. The broker must reconcile multiple intents by performing semantically-aware multi-objective optimization across radio access network (RAN) spectrum, transmission modes, edge computing capacities, and in-network caching placements. The broker's intelligence draws on reinforcement learning policies that have been trained offline to adapt to varying channel states and traffic patterns, while also incorporating online adjustments guided by intent satisfaction measurements [13].

Beneath the broker, cross-domain infrastructure controllers such as the RAN intelligent controller in open radio access network architectures, the 5G core network slicing framework, and multi-access edge computing orchestrators execute the resource assignments [14, 15]. The feedback loop closes through semantic-quality monitoring agents that continuously evaluate whether the deployed configuration preserves the required semantic fidelity. These agents report intent drift back to the interpretation engine, triggering re-evaluation or re-negotiation. The loop thus realizes an intent-driven cognitive network that operates at the semantic layer, continuously reconciling user expectations and physical constraints.

4. LLM-Driven Intent Interpretation and Semantic Mapping

The transition from a user's linguistic expression to a resource allocation directive involves several levels of semantic transformation. LLMs, by virtue of their pre-training on vast corpora, have implicitly acquired relational models of causality, function, and domain constraints. When prompt-engineered appropriately, they can decompose a high-level intent such as "support immersive telepresence for a distributed board meeting with prioritization of

the chairperson's stream" into sub-intents relating to video resolution, haptic feedback latency, prioritization rules, and failure modes. The model leverages its internal representations to associate these sub-intents with known semantic communication primitives such as joint source-channel coding configurations optimized for human perception, task-oriented feature extraction for motion synchronization, and differential treatment of control-plane and data-plane traffic. The system must then ground these associations against a capability registry that enumerates available semantic encoders, their rate–distortion characteristics, and their operational bounds. This grounding step constitutes a critical bridge between the continuous, flexible reasoning of the LLM and the discrete, measurable world of physical infrastructure.

Because the LLM operates over natural language rather than mathematical formulations, the mapping process must incorporate deterministic validation layers that check consistency and feasibility. For instance, if the LLM suggests a semantic encoder that delivers 10-millisecond perception-latency for haptic feedback, the validation layer cross-references this against the edge computing round-trip time budget and channel coherence time, flagging impossibilities. In this manner, the LLM remains a creative yet bounded reasoner. Context-aware retrieval further enhances reliability by supplying recent traffic prediction insights, which can anticipate congestion events and allow the LLM to pre-emptively soften intents or suggest fallback semantic profiles. Recent advances in spatial-temporal traffic prediction using transformer-based models have shown that accurate medium-term forecasts can significantly improve resource provisioning for latency-sensitive applications [9]. Integrating such forecasts into the intent mapping pipeline allows the LLM to reason not only about current intent but also about how network dynamics may erode intent satisfaction in the near future.

The semantic mapping is inherently multi-scale. A single intent may span personal area networks, RAN slices, and core cloud resources simultaneously. The LLM-compiled intent graph must therefore be partitioned into sub-graphs assigned to different orchestration domains while preserving global consistency. Here, the architectural choice between centralized and federated LLM deployments becomes salient. A centralized model can maintain a holistic view but introduces a single point of decision latency and potential privacy exposure. Distributed or edge-native LLMs can respond faster and keep sensitive data localized but may lack the contextual breadth needed for cross-domain coordination. The design space suggests a hybrid model in which lightweight distilled LLMs operate at the edge for localized, real-time intent adjustments, while a more capable, global LLM periodically resolves inter-domain conflicts and updates semantic policies.

5. Resource Allocation Mechanisms and Network Slicing

Resource allocation in semantic-aware networks inherently differs from throughput-centric models because the optimization objective shifts toward maximizing the fulfillment of semantic intents under energy, bandwidth, and computational budgets. Network slicing, as a fundamental enabler in 5G and an anticipated cornerstone of 6G, provides isolated and customizable logical networks over a shared physical infrastructure [14]. In an LLM-driven intent-aware framework, slices are no longer predefined static templates but dynamically instantiated scopes whose properties are derived from the intent graph. For example, a slice supporting a smart grid use case might be configured with ultra-reliable semantic transformers for teleprotection and moderate-throughput semantic codecs for meter data aggregation, all bundled under a single intent comprising safety, efficiency, and latency objectives.

The resource allocation process operates at multiple timescales. At a slow timescale, the broker solves a global resource provisioning problem—assigning spectrum partitions,

computing allocations, and priority classes based on aggregate intents—using models learned offline through multi-agent deep reinforcement learning where agents represent different slice owners negotiating utility [15]. At a fast timescale, per-slot scheduling decisions are made by lightweight, near-real-time controllers that adhere to the slice-specific semantic policies pushed by the broker. The LLM's role is primarily at the slow timescale, where it translates changes in intent into updated slice blueprints and resource budgets. However, when intents conflict, such as during an emergency that requires repurposing resources from entertainment slices to public safety slices, the LLM can mediate a rapid re-evaluation, reasoning about the ethical priorities and generating new resource boundaries that downstream controllers then enforce.

A central tension in resource allocation for semantic flows is the trade-off between semantic precision and resource efficiency. More accurate semantic communication may demand heavier encoder–decoder models that consume greater computational resources and introduce processing delay. The LLM-based intent interpreter must therefore calibrate the semantic fidelity requirements against the overall system state. It achieves this by treating intent expressions not as absolute targets but as utility functions that map semantic quality measures, such as learned perceptual similarity or task accuracy, to satisfaction levels. The resource broker then uses these utility functions to perform joint semantic-resource optimization. The inclusion of predictive traffic intelligence from models capturing spatiotemporal dynamics ensures that resource reservations proactively accommodate predicted load shifts, mitigating the risk of intent violations [9].

Operationalizing such mechanisms in real-world environments requires careful integration with standardized management interfaces. The O-RAN architecture, with its non-real-time and near-real-time RAN intelligent controllers, provides a natural hosting environment for the broker and the LLM interpretation engine, while 3GPP management and orchestration functions handle slice lifecycle operations [16, 17]. This integration demands well-defined application programming interfaces that expose resource capabilities and semantic monitoring metrics in a language that a prompted LLM can reason over, potentially through structured text-based representations of network state. The performance and reliability of this interface chain directly influence the overall intent satisfaction rate, making it a critical system design consideration.

6. Governance, Trust, and Sustainability Considerations

Placing an LLM at the heart of resource allocation raises multifaceted governance challenges that go beyond technical performance. Trustworthiness is paramount; an LLM-derived resource policy must be explainable to both human operators and regulatory bodies. Without explainability, incidents such as unfair resource denial or unintended priority inversions cannot be properly audited or rectified. Contemporary LLMs can generate fluent explanations, but their fidelity to actual decision logic is not guaranteed, a problem known as post-hoc rationalization. Thus, the architecture must incorporate structured explanation modules that trace the intent-to-policy pathway through the intent graph and validation layers, rather than relying solely on the LLM's narrative output.

Fairness across users, verticals, and services becomes a pressing concern when natural language intent can inadvertently encode biased priorities. The LLM, having been trained on socio-linguistic patterns, may reinforce hidden biases unless explicitly constrained by fairness guardrails. For instance, an intent phrased in technical jargon by a sophisticated enterprise may receive more precise resource mappings than an equivalently intended but colloquially

expressed request from a small community network. Governance frameworks must therefore standardize intent expression protocols that level the playing field, possibly adopting controlled natural language subsets, and must continuously monitor resource allocation outcomes for disparate impact.

The sustainability implications of large models are equally significant. Training and inference of state-of-the-art LLMs consume substantial energy, which directly conflicts with the energy-efficiency goals central to 6G design [3, 13]. A naive deployment that routes every intent interpretation request to a massive cloud-hosted LLM would be ecologically and economically unsustainable. The system must exploit model cascades, where small, fine-tuned edge models handle routine intents and escalate only ambiguous or novel cases to larger models, thereby trading off cognitive capability against energy expenditure. Further, model distillation, quantization, and sparsity techniques tailored for intent interpretation can reduce the inference footprint without unduly sacrificing performance. Regulators and standards bodies such as the International Telecommunication Union are beginning to define assessment frameworks for the environmental sustainability of AI-enhanced networks, and the proposed architecture must pro-actively align with these emerging guidelines [18].

7. Trade-offs and Deployment Challenges

The operational deployment of LLM-driven intent-aware semantic resource management confronts several irreducible trade-offs. The first is the latency versus richness trade-off. Fully expressive LLM inference can take several seconds, far exceeding the tens-of-milliseconds budget of fast-timescale resource actions. Partitioning the problem into slow intent interpretation and fast enforcement, as described earlier, mitigates but does not eliminate this tension: whenever an urgent intent change occurs, the end-to-end reaction latency includes the LLM inference plus policy translation and distribution delay. Feasibility studies suggest that for many 6G use cases, including those in industrial automation and telemedicine, such latency may be acceptable if the system resorts to pre-computed fallback policies during the interpretation window, but the performance guarantees require rigorous worst-case analysis.

The second trade-off involves model generality versus domain specificity. General-purpose LLMs possess broad reasoning capabilities but are expensive to fine-tune and may exhibit unpredictable behavior in highly technical networking contexts. Conversely, domain-adapted models achieve better accuracy and efficiency but demand continuous retraining as network capabilities and semantic coding libraries evolve. A hybrid approach, combining a general foundation model for intent parsing with specialized adapter modules for semantic mapping, offers a promising compromise. However, the maintenance burden of adapter ecosystems and the potential for version skew between foundational and adapter models introduce lifecycle management complexity.

Privacy and data sovereignty present further challenges. User intents, especially when enriched with environmental context, may inadvertently reveal sensitive patterns about enterprise operations or individual behaviors. Centralizing intent processing in a public cloud LLM service would expose such information to third parties and conflict with data localization regulations. Federated architectures that process intents at the edge or within private operator clouds, and only share anonymized, aggregated knowledge for global model improvement, help reconcile functionality with privacy [19]. Simultaneously, adversarial intent injection must be considered as a new attack surface; maliciously crafted intents could trick the LLM into allocating excessive resources or downgrading security mechanisms.

Robust intent sanitization layers and continuous anomaly detection on intent patterns emerge as necessary defenses.

Interoperability across multiple vendors and technology generations adds further friction. In a multi-vendor 6G landscape, the intent interpretation engine must translate intents into policies executable on heterogeneous equipment, each with its own semantic encoders and resource abstraction models. Standardizing the semantic resource broker interface and the semantic capability description language would reduce integration cost, but such standardization is still in its infancy. Legacy 5G devices and network functions that lack semantic communication stacks will co-exist, and the system must gracefully degrade intents into traditional quality-of-service parameters when semantic operation is not possible, a translation task that itself requires careful LLM prompting and validation.

8. Future Perspectives and Research Directions

Looking forward, the integration of LLMs with 6G semantic communication networks will evolve along several dimensions. Multi-modal LLMs that jointly process text, images, radio-frequency maps, and sensor readings hold the potential to elevate intent interpretation to an even more contextually aware plane, where a user's gesture, spoken command, and device status all contribute to a unified intent representation. This will necessitate research into cross-modal semantic grounding and resource allocation that respects diverse modality-specific distortion measures.

Federated and continual learning will become essential as networks accumulate intent-resolution traces. Instead of periodically retraining a central LLM, which is costly and risks staleness, a federation of operator-edge models could collaboratively refine the intent mapping under strict privacy constraints, while a lightweight orchestration unit aggregates learning without exposing raw intent data. The challenge of non-stationary intent distributions must be addressed, as socio-economic trends, regulatory shifts, and new vertical industries continuously reshape the vocabulary and structure of intents.

Digital twin platforms are poised to play a pivotal role by providing a high-fidelity simulated replica of the physical network where LLM-generated policies can be tested for safety, fairness, and efficiency before deployment. Through a continuous integration and delivery pipeline for network intelligence, a proposed policy could be stress-tested against thousands of scenarios synthesized from historical traffic patterns and adversarial intent stress-tests, reducing the risk of policy-induced outages. Human-in-the-loop governance mechanisms need to be cultivated whereby domain experts can override or refine LLM outputs in critical situations, and these corrections are fed back as ground truth for future learning cycles.

Finally, economic models of intent-driven resource allocation deserve scholarly attention. When multiple tenants express overlapping intents, an economic framework that prices semantic resource units—accounting for their scarcity and the value they deliver to each tenant—can resolve conflicts in a transparent and incentive-compatible manner. Such a market would require the LLM, or an adjacent agent, to understand value propositions and negotiate intents on behalf of users, opening a frontier at the intersection of economics, AI, and network engineering. Standardization bodies, including 3GPP and the O-RAN Alliance, will need to define intent expression grammars, semantic capability descriptors, and policy enforcement points that collectively enable this ecosystem [20, 21].

9. Conclusion

This paper has examined the system-level architecture, design tensions, and governance imperatives of large language model-driven intent-aware resource allocation for 6G semantic communication networks. By positioning the LLM as an intent interpreter and semantic mapper, we have outlined a cognitive control loop that translates natural language user goals into semantically coherent resource assignments across a distributed infrastructure. We argued that such a framework must be understood not as a monolithic AI application but as a carefully orchestrated assembly of intent dialogue, retrieval-augmented reasoning, validation, resource brokering, and monitoring, all operating within a multi-timescale governance structure. The analysis highlighted critical trade-offs between inference latency and scheduling timeliness, between model generality and domain fidelity, and between central intelligence and edge autonomy. Sustainability and trust considerations demand that energy-conscious model cascades, fairness-aware guardrails, and explainable auditing pathways be embedded from the earliest design stages. Future efforts should focus on multi-modal intent acquisition, federated continual adaptation, digital twin-based policy validation, and the design of economic mechanisms that reconcile competing intents in a multi-stakeholder environment. As 6G transitions from vision to realization, the responsible integration of generative AI into network management will not only enhance efficiency but also redefine the very nature of human–infrastructure interaction, placing meaningful, trustworthy communication at the core of the next-generation digital society.

References

1. Saad, W., Bennis, M., & Chen, M. (2020). A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. IEEE Network, 34(3), 134–142.

2. Letaief, K. B., Chen, W., Shi, Y., Zhang, J., & Zhang, Y. J. (2019). The roadmap to 6G: AI as a main driver for future wireless networks. IEEE Communications Magazine, 57(8), 78–84.

3. Strinati, E. C., Barbarossa, S., Gonzalez-Jimenez, J. L., Kténas, D., Cassiau, N., Maret, L., & Dehos, C. (2019). 6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication. IEEE Vehicular Technology Magazine, 14(3), 42–50.

4. Popovski, P., Simeone, O., Boccardi, F., Gündüz, D., & Sahin, O. (2020). Semantic-effectiveness filtering and control for post-5G wireless communication. IEEE Journal on Selected Areas in Communications, 38(5), 1015–1027.

5. Xie, H., Qin, Z., Li, G. Y., & Juang, B. H. (2021). Deep learning enabled semantic communication systems. IEEE Transactions on Signal Processing, 69, 2666–2679.

6. Gündüz, D., de Kerret, P., Sidiropoulos, N. D., Gesbert, D., Murthy, C. R., & van der Schaar, M. (2020). Machine learning in the air. IEEE Journal on Selected Areas in Communications, 37(10), 2184–2199.

7. IMT-2030 (6G) Promotion Group. (2021). White paper on 6G vision and requirements. China Academy of Information and Communications Technology.

8. Jiang, W., Han, B., Habibi, M. A., & Schotten, H. D. (2021). The road towards 6G: Opportunities, challenges, and the road ahead. IEEE Open Journal of the Communications Society, 2, 756–785.

9. 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.

10. Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681–694.

11. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., … Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

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. Mao, B., Tang, F., Kawamoto, Y., & Kato, N. (2022). AI models for green communications towards 6G. IEEE Communications Surveys & Tutorials, 24(1), 210–247.

14. Habibi, M. A., Nasimi, M., Han, B., & Schotten, H. D. (2019). A comprehensive survey of RAN architectures toward 5G mobile communication system. IEEE Access, 7, 70371–70421.

15. Chen, M., Gündüz, D., Huang, K., Saad, W., & Poor, H. V. (2022). Resource allocation for wireless networks: An optimization and machine learning perspective. Proceedings of the IEEE, 110(6), 922–951.

16. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

17. 3GPP. (2023). System architecture for the 5G System (5GS) (TS 23.501, Release 18). 3rd Generation Partnership Project.

18. ITU-T. (2020). Framework for evaluating intelligence levels of future networks including IMT-2020 (Recommendation Y.3173). International Telecommunication Union.

19. Taleb, T., Samdanis, K., Mada, B., Flinck, H., Dutta, S., & Sabella, D. (2017). On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration. IEEE Communications Surveys & Tutorials, 19(3), 1657–1681.

20. Zhou, Y., Liu, L., Wang, L., Hui, N., Cui, X., Wu, J., … Hanzo, L. (2020). Service-aware 6G: An intelligent and open network based on convergence of communication, computing, caching, and control. IEEE Journal on Selected Areas in Communications, 38(7), 1470–1487.

21. Dohler, M., Heath, R. W., Lozano, A., Papadias, C. B., & Valenzuela, R. A. (2011). Is the PHY layer dead? IEEE Communications Magazine, 49(4), 159–165.

22. Saad, W., Durrani, S., & Bennis, M. (2024). Generative AI for physical layer communications: A survey. IEEE Communications Surveys & Tutorials, early access.

23. Behringer, M., Pritikin, M., Bjarnarson, S., Clemm, A., Carpenter, B., Jiang, S., & Ciavaglia, L. (2015). Autonomic networking: Definitions and design goals (RFC 7575). Internet Engineering Task Force.

24. Shao, Y., Ozfatura, E., Perotti, A., Popovski, P., & Gündüz, D. (2022). Task-oriented communication for edge inference. IEEE Journal on Selected Areas in Communications, 40(12), 3503–3518.