Proceedings of CONF-SPML 2027 Symposium: Intelligent Network Security and Machine Learning for Communication Systems
DOI: 10.54254/2755-2721/2027.MELB38733

Towards a Universal Semantic Basis: A Survey on the Fundamental Units of Semantic Communication in the LLM Era

Kaiyi Chen

Department of Engineering, King's College London, London, UK
K23051815@kcl.ac.uk

Abstract. As communication technology enters the 6G era, traditional "bit-based" syntax-level (Level A) communication is hitting physical bottlenecks from the Shannon limit. In recent years, the rise of Large Language Models (LLMs), especially technologies aimed at distributed deployment and prior knowledge sharing, has made LLMs a strong candidate for shifting human communication to "intent and semantics" (Level B) transmission. Moving semantic communication toward engineering practice and standardisation requires a unit equivalent to "bits" in traditional systems. However, most current studies on LLM collaborative communication focus on edge-cloud synchronisation of intermediate state tensors like the KV Cache. This article introduces the structural characteristics of the KV Cache and shows why it cannot serve as the minimum measurement unit for semantic communication. Meanwhile, building on existing intelligent agent collaboration technologies, this article derives the basic properties the fundamental unit of pan-semantic communication must possess. The analysis shows that intermediate tensors such as the KV Cache are constrained by strong model coupling and dimensional expansion; they serve only as temporary computational residues rather than true semantic carriers. Therefore, this paper concludes that a valid minimal unit of semantic communication must be model-agnostic and topologically invariant. Defining such a unified, intent-based unit helps communication systems move from rigid state synchronisation to flexible semantic routing in future 6G networks.

Keywords: Semantic Communication, LLMs, Shannon Limit, KV Cache, Intelligent Communication

1. Introduction

Since Claude Shannon laid the foundation of information theory in 1948, modern communication systems have striven over the past few decades to approach the physical limit curve of channel capacity widely known as the Shannon limit. From Turbo codes, which were the first coding scheme that approached this limit and also estimated the 3G/4G era [1], to Polar codes in the 5G era [2], which have been strictly proven mathematically to reach this limit, the encoding gain at the syntax level (Level A, which precisely transmits the underlying binary bits) has been almost squeezed out.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.MELB38733

With 6 G networks' increasingly stringent requirements for massive connections and extremely low latency, relying solely on bandwidth expansion or improving signal-to-noise ratio (SNR) is no longer enough to bridge the physical divide. The bottom-up transition of the communication paradigm is imminent.

As early as 1949, Weaver proposed three conceptual levels of communication: the syntax level (Level A) focusing on the accurate transmission of signals, the semantic level (Level B) addressing how accurately the transmitted signal expresses the transmitter's intention, and the effectiveness level (Level C) concerning whether the received information elicits the expected reaction [3]. This framework guides future communication networks toward semantic communication. However, semantic communication has long remained theoretical because no universal tool can deeply represent massive prior knowledge. In recent years, the explosive development of Large Language Models (LLMs) has filled this gap. Through massive data pre-training, LLMs naturally possess powerful feature extraction and intent-parsing capabilities grounded in the world-knowledge distribution. By deploying LLMs at both the transmitting and receiving ends, a communication system no longer relies on bit-by-bit replication of raw data. Instead, it becomes an interaction of highly abstract semantic features, offering robustness and coding gains at extremely low signal-to-noise ratios that traditional systems cannot match.

With the distributed deployment of LLMs in edge networks, such as edge-cloud collaborative architectures based on ISCC, efficiently transmitting the intermediate states of large models has become a research hotspot. Currently, the mainstream trend in academia focuses on the extreme compression and transmission of the KV Cache within attention mechanisms. Yet transmitting the KV Cache is basically a process of computing power offloading and state synchronisation in distributed computing. As an intermediate state tensor in autoregressive decoding, it has two major flaws. First, the model is strongly coupled: its tensor dimensions and structure are tightly bound to specific neural networks, so heterogeneous models cannot generalise or parse it. Second, it suffers from dimensional expansion, as the KV Cache size grows linearly or even quadratically with context length. Because this computing residue expands over time and lacks invariance, it cannot work as the core unit of communication.

To find a minimum semantic unit with the same status as the "bit" in Level A systems, this paper derives the properties of the basic unit applicable to general semantic communication. The rest of this article first presents a theoretical falsification, showing from the viewpoints of model generalisation and dimensional convergence why the KV Cache and similar intermediate tensors cannot serve as basic semantic units. Next, define a model-independent basis for semantic communication units drawing on high-dimensional topological invariance and generalised semantic entropy. Lastly, we discuss how transmitting this new measurement unit fits into 6G ISCC networks to support future edge intelligence.

The main value of this research is updating the theoretical path for next-generation communication systems. For the communication industry, defining a universal semantic unit helps overcome the hardware and bandwidth limits of traditional bit transmission, enabling low-power edge devices to participate in complex, intent-based interactions in 6G IoT applications. For future researchers working on related topics, this study shifts the focus away from hardware-tied state synchronisation toward a mathematically invariant metric for semantic communication, offering a solid reference for designing future cross-modal and heterogeneous semantic networks.

DOI: 10.54254/2755-2721/2027.MELB38733

2. Past explorations in LLM semantic communication and the limitations

2.1. Discovery of LLM' semantic gain

The academic community has preliminarily verified that LLMs can serve as semantic transceivers. The LLM-based semantic communication (LLM-SC) system proposed by Wang et al. shows that the system can achieve zero-error semantic-level transmission when the signal-to-noise ratio (SNR) is greater than 3 dB. Compared with traditional communication systems based on Shannon theory, LLM-SC achieves a coding gain of about 8 dB at a bit error rate of 10^{-3} without adding any traditional channel error-correction coding [4]. This result greatly reduces computation and bandwidth pressure on the radio frequency (RF) front end. It also shows that sharing a large model's pre-trained knowledge base between the transmitter and receiver can resist channel noise at the physical level, a solid step toward Level B human communication.

2.2. The rise of KV Cache transmissions

As LLMs extend to edge computing devices, the communication bottleneck in distributed inference has become increasingly apparent. To reduce computational pressure on the edge, researchers have begun exploring direct state transfer between LLMs. An early example is the prompt caching technology used by Anthropic. Instead of discarding the Key-Value (KV) matrices formed in previous model calculations, this method saves them. When the input prefix of subsequent requests stays the same, these KV matrices are directly loaded back into the model [5]. This confirms that long-term storage, reuse, and transmission of hidden layer states in large models are practical.

This directly gave rise to subsequent communication protocols based on the KV Cache. For example, Kriuk and Ng focused on "Cache-to-Cache" communication and introduced a protocol called Q-KVComm [6]. This protocol transmits the KV Cache directly through adaptive compression technology and reaches a compression ratio of 5 to 6 times. Experiments show that this method keeps LLM performance stable across scales while maintaining high semantic fidelity and coherence. Right now, this approach is the mainstream technical route for collaborative communication between cloud and edge models.

2.3. Theoretical limitations of the KV Cache as a semantic unit

However, from an information-theoretic viewpoint, an overlooked theoretical blind spot emerges: compressing and transmitting the KV Cache is essentially "computational state synchronisation" under a distributed inference framework, not true semantic communication. Defining the KV Cache as the underlying unit of semantic communication in large models runs into three theoretical limits that cannot be bypassed.

The first limit is strong model coupling. The KV Cache is an intermediate state tensor tightly bound to a specific Transformer architecture. If a cloud sender uses a dense model with billions of parameters, while the edge receiver runs a heavily pruned or quantized Mixture-of-Experts (MoE) model, the two models are completely different in parameter count, attention head count, and expert routing mechanisms. As a result, the KV cache the edge device receives becomes meaningless floating-point numbers. A true basic unit of communication must have model-agnostic properties so heterogeneous networks can work together.

The second limit is dimensional expansion. In classical information theory, "bits" have atomic, fixed dimensional properties. By contrast, the physical volume of the KV Cache grows linearly or

DOI: 10.54254/2755-2721/2027.MELB38733

quadratically as the context sequence length increases. In long-context communication, using such a time-varying variable undermines the convergence and compactness that basic communication units should have.

The third limit is that the KV Cache acts as a computing residue rather than a basis for intent. Inside the system, the KV Cache is basically managed as virtual memory pages. It is a historical cache kept only to avoid repeated calculations during autoregressive decoding, rather than a condensed form of "semantic entropy" from the original information. It carries the hardware-layer computing-logic state instead of pure communication intent.

This means that as long as semantic communication is restricted to specific computing tensors, building a unified pan-semantic communication network is impossible.

3. Deduction of the basic unit properties of semantic communication: units proposed based on different studies

The concept of "named sentence similarity," introduced by Xie et al., represents an early attempt to benchmark semantic transmission error rates in related research. This method evaluates the completion of information transmission by calculating the cosine similarity between two semantic vectors, expressed as [7]:

$$\mathrm{match}(\hat{s}, s) = \frac{B_\Phi(s) \bullet B_\Phi(\hat{s})^T}{\|B_\Phi(s)\| B_\Phi(\hat{s})\|\|} \tag{1}$$

Where s denotes the original transmitted sentence, $\hat{s}$ represents the received sentence recovered at the destination, and B_ϕ signifies the semantic feature extraction model BERT [8]. However, as this research predated the widespread integration of Large Language Models (LLMs), it primarily focused on sentence-level similarity rather than exploring an underlying, indivisible fundamental unit for semantic communication.

To formally define the equivalents of "Shannon entropy" and "bits" in the semantic domain, two revolutionary concepts emerged: the Semantic Base (SEB) proposed by Zhang et al. from the Beijing University of Posts and Telecommunications, and Semantic Entropy proposed by the OATML group at Oxford University.

Specifically, SEB is a comprehensive and highly abstract representation model that extracts multimodal features of information through specific knowledge graphs or other organised expressions, and then constructs a semantic representation framework at the macro level [9].

The OATML team went further; following Shannon's approach of measuring symbol sequence disorder, they extended this logic to measure intent uncertainty. The team extended the standard Shannon entropy formula, formulating the equation for ideal semantic entropy [10]:

$$SE(x) = -\sum_c p(cx) \log p(cx) = -\sum_c((\sum_{s\epsilon c} p(sx))\log[\sum_{s\epsilon c} p(s|x)]) \tag{2}$$

where x represents the given context or input, s denotes a specific generated token sequence, and c represents a semantic equivalence class encompassing all generated sequences that share the exact same logical meaning. Crucially, the scope of SE(x) is no longer isolated word sequences, but the underlying meaning distribution. This mathematically proves that even if the token-level syntax entropy is extremely high due to varied wording, the semantic entropy approaches zero as long as the outputs are logically equivalent.

DOI: 10.54254/2755-2721/2027.MELB38733

Drawing upon these theoretical foundations, we deduce that the Universal Semantic Unit (USU) must possess three core properties. First, it must act as a "Minimal Tensor Space": a highly pure intent coordinate system that maximises source information compression while maintaining the target semantic entropy. As long as both transceivers share the basic prior knowledge distribution, lossless mapping and restoration of intent can be achieved within this universal coordinate system. Second, it must exhibit "Topological Invariance": derived from the discrete and hierarchical structures of natural language, this invariant topological structure ensures that the unit can achieve zero-error transmission even under extremely low signal-to-noise ratios. Finally, it must guarantee "Dimensional Convergence": unlike the KV Cache, which expands indefinitely with sequence length, the dimension of the semantic communication basic unit should be strictly correlated with the intrinsic "semantic entropy" of the source information. For a lengthy text with low core intent(semantic entropy), the USU will highly collapse into a low-order tensor, effectively resolving the dimensional dilation problem inherent in long-context communication.

4. Meaning for edge intelligence and ISCC

A unified universal semantic unit will be highly valuable for future 6G networks and the Internet of Things (IoT). Edge IoT devices face strict power and computing limits. Even with KV Cache transmission, the edge side still needs to deploy lightweight large models with the same architecture to handle the heavy state alignment of attention mechanisms.

After obtaining a universal semantic unit, edge nodes do not need to run complete LLMs or keep large intermediate states. This allows us to deploy very lightweight "semantic feature mapping networks" on the edge side, which are only responsible for converting environmental intentions into low-dimensional units. This approach separates transmission from computing power, allowing edge devices that originally couldn't support distributed inference for large models to access cloud-based knowledge bases over ultra-low bandwidth.

Moreover, based on the SEB concept, the universal semantic unit can compress not only language text but also multimodal perceptual data. Both the images from edge cameras and the waveforms from RF sensors can be treated as forms of "environmental language." Once the core "semantic entropy" is extracted, it can be mapped onto a unified semantic base for broadcasting. This greatly reduces the difficulty of combining sensing, computing, and communication, enabling true fusion of these three processes under a unified measurement unit.

At the network level, a unified semantic unit can achieve adaptive transmission in 6G networks. Inside the network, routers will no longer perform simple bit-packet forwarding; instead, they will act as intelligent routers based on upstream and downstream semantic transmission. They can dynamically adjust the transmission resolution and dimension based on channel quality. This means that in extreme cases where the channel condition is very poor, the network can still ensure the delivery of the core logical intent.

In short, developing the universal semantic unit offers an effective way to solve the bottleneck of 6G edge distributed inference.

5. Conclusion

Through a comprehensive theoretical analysis of LLM-driven communication paradigms, this study finds that current mainstream distributed inference techniques—such as KV Cache-based state synchronization—are fundamentally restricted by strong model coupling and dimensional expansion. They are essentially mechanical transfers of underlying computing states and cannot

DOI: 10.54254/2755-2721/2027.MELB38733

serve as the foundation for pan-semantic communication. Furthermore, our research reveals that genuine semantic measurement units must be entirely decoupled from specific neural network architectures and autoregressive computation residues. By analysing previous definitions represented by "Semantic Entropy," this study successfully derives the essential properties required for future universal semantic units. Establishing a semantic measurement unit based on these properties not only theoretically resolves the dimensional disaster associated with long-context transmission but also provides a unified measurement standard for 6G edge networks.

However, as a forward-looking review aimed at reconstructing the communication paradigm, this article primarily focuses on establishing conceptual boundaries and system architectures from an information-theoretic perspective, which inevitably leaves limitations in strict mathematical derivation. Although the abstract features of this unit are defined based on semantic entropy and high-dimensional tensors, a rigorous proof of its tensor projection in a specific topological space and a quantitative deduction of its dimensional convergence at the calculus level have yet to be provided.

Recognising that the leap from a theoretical proposal to a complete axiomatic system cannot be exhausted by a single review, this article aims to stimulate further research and calls for the introduction of rigorous pure mathematical analysis into this field. Ultimately, discovering and refining the strict mathematical boundaries of universal semantic basic units will be the core proposition for the next generation of communication theory and the distributed evolution of artificial general intelligence (AGI).

References

[1] Berrou, C., Glavieux, A., & Thitimajshima, P. (1993). Near Shannon limit error-correcting coding and decoding: Turbo-codes. 1. In Proceedings of ICC '93 - IEEE International Conference on Communications (pp. 1064–1070). https://doi.org/10.1109/ICC.1993.397441

[2] Arikan, E. (2009). Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. IEEE Transactions on Information Theory, Vol(Issue), Pages. https://doi.org/10.1109/TIT.2009.2021379

[3] Weaver, W. (2017). Recent contributions to the mathematical theory of communication. ETC: a review of general semantics, 74(1/2), 136-157.

[4] Wang, Z., Zou, L., Wei, S., Li, K., Liao, F., Mi, H., & Lai, R. (2024). Large-language-model-enabled semantic communication systems. arXiv preprint arXiv:2407.14112.https://doi.org/10.48550/arXiv.2407.14112K. Elissa, "Title of paper if known," unpublished.

[5] Anthropic. (n.d.). Prompt caching. Anthropic API Documentation. Retrieved August 30, 2026, from https://docs.anthropic.com/en/docs/build-with-claude/prompt-caching

[6] Kriuk, B., & Ng, L. (2026). Q-KVComm: Efficient multi-agent communication via adaptive KV cache compression. In *2026 Second International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)* (pp. 15–21). https://doi.org/10.1109/ICMSCI67830.2026.11469367

[7] Xie, H., Qin, Z., Li, G. Y., & Juang, B.-H. (2020). Deep learning-enabled semantic communication systems. arXiv. https://doi.org/10.48550/arXiv.2006.10685

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[9] Zhang, P., Xu, W., Gao, H., Niu, K., Xu, X., Qin, X., Yuan, C., Qin, Z., Zhao, H., Wei, J., & Zhang, F. (2021). Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks. *Engineering, 8*. https://doi.org/10.1016/j.eng.2021.11.003

[10] Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. arXiv. https://doi.org/10.48550/arXiv.2302.09664