ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2025.1834.393

NEUROMORPHIC-INSPIRED HYBRID COGNITIVE MODEL FOR SELF-OPTIMIZING RESOURCE MANAGEMENT IN 6G EDGE NETWORKS

Reji K Kollinal[1*], Mariya T Cheeran[2]

*[*1]Assistant Professor, BPC College, Piravom, India. e-mail: rejibpc@gmail.com, orcid: https://orcid.org/0000-0001-5757-2411*
[2]Senior Manager, Reply India, Kochi, India. E-mail: mariyatcheeran@yahoo.com, orcid: https://orcid.org/0009-0009-4205-3235

Received: September 01, 2025; Revised: October 17, 2025; Accepted: November 24, 2025; Published: December 30, 2025

SUMMARY

Introduction: The 6G world requires connected intelligence, but there is a crucial paradox between the standards of Large Language Model (LLM) and edge constraints. The premium devices have up to 6-12GB of DRAM, whereas the typical 175B models need 350GB of storage, which is 30 times that of the premium version. Literature Survey: It has been proposed that the bandwidth can be reduced by 90 % with Semantic Communication (Scom) and Edge Semantic Cognitive Intelligence (ESCI). Besides, neuromorphic-based Spiking Neural Networks (SNNs)-model quantization (INT4/INT8) are also known to be necessary to achieve order-of-magnitude energy efficiency (J/token) on resource-constrained hardware. Methodology: This paper proposes a Hybrid Cognitive Model utilizing a three-tier Cloud-Edge-Device hierarchy. The model integrates event-driven neuromorphic principles with self-optimizing resource management, utilizing paged KV-cache and resource-aware agents for dynamic task offloading. Results: Quantitative evidence is used to show that the hybrid strategy helps to address the 30x resource gap by attaining a 10-100x energy-per-token efficiency due to event-driven neuromorphic sparsity. Statistical analysis makes it evident that semantic filtering substantially reduces communication overhead and maintains reasoning faithfulness by 90 %, and, effectively, it keeps the thermal conditions of devices stable in the case of prolonged 6G edge communications. This model can be used to make sustainable and multi-step thinking on the edge. The hybrid solution achieves the 6G vision of pervasive intelligence by bridging the hardware-software gap via cross-layer co-design.

Key Words: *6G Edge Intelligence, Neuromorphic Computing, Semantic Communication, Large Language Models (LLMs), Spiking Neural Networks (SNNs), Resource-Aware Agents, Cognitive Edge Computing*

INTRODUCTION

The sixth-generation (6G) wireless networks introduce a radical paradigm shift in the traditional data-based communication model of connected intelligence [1]. The new era brings with it the network playing a more important role, beyond the bit-level transmission; it becomes a distributed cognitive fabric with the Edge Intelligence (EI) giving the autonomy to make decisions and reasoning in a human manner at the edge of the network [2][3]. It is hoped that 6G will deliver ubiquitous, low-latency, and

privacy-preserving intelligence by integrating edge computing with state-of-the-art artificial intelligence (AI) [5][13]. This is aimed at leaving behind the narrow perception tasks in favour of Edge Semantic Cognitive Intelligence (ESCI), where networks are able to reason, act, and understand the meaning of data in real-time [4].

The actualization of this vision is experiencing a terminal deployment contradiction. The backbone of the modern cognitive activity, Large Language Models (LLMs) and autonomous agents, demands massively in terms of computational and memory needs, which are fundamentally antagonistic to edge hardware [10]. Quantitative analysis indicates the existence of a huge disparity: a typical high-end edge device, including smartphones and IoT gateways, has no more than 6-12GB of DRAM, and a typical 175B parameter model needs more than 350GB of storage [15]. This is a resource deficit of more than 30x, made worse by the tight power and thermal requirements of mobile hardware, which is usually in a sub-10W power budget.

Neuromorphic computing is an innovation that can fill this gap. Being a brain-inspired system, neuromorphic systems combine event-driven computation and spiking to obtain orders of magnitude of energy efficiency advantages over conventional von Neumann systems [11]. These systems greatly minimize the amount of power used by continuous inference by processing information only when an event or a spike happens. With neuromorphic principles incorporated, edge agents can enable high-performance cognitive functions that do not drain local battery resources or cause thermal throttling [18][19].

The paper aims to optimize resources provided self-optimally in a self-managed 6G edge network. The model coordinates deep learning systems using semantic communication guidelines to sift through a lot of irrelevant data and emphasize the meaning, which minimizes upstream bandwidth by as much as 90 %. The proposed system combines the event-based neuromorphic sparsity with adaptable resource-aware agents so that it can be both sustainable and perform multi-step reasoning at the edge. This hybrid methodology addresses the hardware-software gap by using cross-layer co-design, which satisfies the 6G need of pervasive and autonomous intelligence.

This paper will be divided into the following structure: Section II will be devoted to the literature survey of the history of semantic information theory, and will examine the existing trends of neuromorphic and cognitive edge computing. The III provides the description of the suggested system architecture, which figures out the three-layer Cloud-Edge-Device structure and the ESCI framework. Section IV explains the self-optimizing resource management mechanisms, which are resource-conscious decision-making, semantic optimization, and run-time memory coordination. Section V discusses representative 6G applications that are intelligent infrastructure maintenance and latency-critical vehicular cognition. Section VI assesses the model performance, addresses normalised measurement protocols, and presents open issues like security and hardware-conscious compilation. In Section VII, the paper ends with a conclusion as to the findings and the future research directions.

LITERATURE SURVEY

Evolution of Semantic Information Theory

The classical communication paradigm, which is based on the information theory of Shannon, is more concerned with the safe transference of symbols irrespective of the meaning contained in them. Nevertheless, the 6G era requires a transition towards Semantic Communication (SCom) that puts an emphasis on the meaning or intent of the message. This theoretical development is traced back to the work of Carnap and Bar-Hillel, who also provided semantic information on the basis of logical probability, differentiating it, however, from the purely statistical perspectives. The modern advancements have accentuated the goal-based frameworks in which communication is concerned with task-oriented goals. The encoding of the intended meaning and the semantic filtering of the transmitted data discard the irrelevant data at the transmitter, eliminating resources and communication latency significantly, and a bandwidth saving of up to 90 % is common.

Edge Cognitive Paradigms

The literature has drawn a clear difference between the traditional Edge AI and the new Cognitive Edge Computing [14][16]. When traditional edge AI performs on a small perception problem, say, image classification or simple object detection, Cognitive Edge Computing is aimed at giving answers to a complex, multi-modal problem, and autonomous decisions [20]. The paradigm stresses maintaining complex cognitive abilities, such as the ability to be aware of the context and reason in many steps, with the extreme limitations of the network edge. The transition is a shift towards agents that do not merely sense their environment but actively reason and plan using it, and a basic reevaluation of the distribution of intelligence throughout the network is necessary [6][7].

Neuromorphic Computing Trends

It is also found that brain-inspired computing is a crucial way forward to sustainable 6G intelligence, especially in power-constrained conditions [8][9][12]. SNNs make use of event-driven processing in which the computation is only performed when a spike is activated, which is extremely energy-efficient as compared to the continuous processing of standard neural networks. Meanwhile, hardware accelerators, e.g., the compute-in-memory (CIM) architectures, are too important to be ignored as it offers 10-100 improvements in energy efficiency by reducing the amount of data that needs to be moved between memory and processors. Recent work is actively investigating how to combine these principles of neuromorphic engineering with Transformer-based models to achieve high-fidelity event-driven inference at the edge.

Resource Management Strategies

The implementation of large-scale models on the edge demands the coordinated optimization of the various layers of the system. Optimization methods: Model optimization. The problem of parameter optimization is important in that it can be used to implement large sets of parameters in the small DRAM of edge devices. Orchestration strategies such as model partitioning and elastic offloading can be used on the system side to partition layers between devices and edge servers to trade off between latency and privacy. Moreover, new innovations like paged KV-cache and hierarchical sparse attention have been introduced to control the memory fragmentation involved in long context reasoning in a multi-turn task of thought [17].

Identification of Research Gaps

Although it has made a great progress, there are still some major gaps in the existing body of research. The majority of the existing frameworks emphasize the accuracy of the perceptions instead of the multi-step reasoning maintenance when compressing the aggressive model. The lack of uniform energy reporting procedures is also significant, and thus prevents the comparative evaluation of sustainable edge AI models. Moreover, the security consequences of low-bit implementation are under-investigated; in particular, quantization-conscious attacks, including bit-flip or fault-injection, do threaten edge-based reasoning integrity in a special way. Lastly, it has no detailed standards of multi-agent collaborative intelligence and modality-aware reasoning in dynamic 6G environments.

SYSTEM ARCHITECTURE FOR COGNITIVE EDGE NETWORKS

Three-Tier Computing Hierarchy

The suggested system is a system that works in a three-layer orchestration that is supposed to facilitate the gap between the model size and the edge restrictions. The Cloud Tier is the knowledge hub world, which is trained on huge amounts of data and has full-scale models. At 6G base stations, there is the Edge Tier, which is a regional coordinator with 10s-100s GB of memory, helping to perform model sharding and low-latency semantic processing. The Device Tier makes direct local inferences in a Watt-scale power constraint and 6-12 GB DRAM constraint. This hierarchy provides an adaptable workload movement, in which thinking jobs are allotted according to the real-time hardware health and latency demands.

Edge Semantic Cognitive Intelligence (ESCI) Framework

The ESCI framework reformulates communication in the sense that it does not focus on bit-level accuracy but semantic-level fidelity. It makes use of an attention-driven sampling process to derive the value of information from raw data. The attention weights (W_a) of the input features model mathematically the Semantic Information Gain (I_s) :

$$I_s = \sum W_a^{(i)} \cdot \log\left(\frac{1}{P(x_i \mid S)}\right) \qquad (1)$$

From Equation (1) $P(x_i \mid S)$is the probability of feature $x_i$being relevant to task S. By discarding non-essential features, the framework generates a "Semantic Bitstream" that reduces upstream bandwidth by up to 90%, directly mitigating the resource occupation at the edge.

Hybrid Neuromorphic-LLM Integration

To enable multi-step reasoning under tight energy budgets, the architecture integrates event-driven Neuromorphic Principles with Transformer architectures. Standard dense layers are replaced with Spiking Neural Network (SNN) blocks that operate on a binary basis. Computation is governed by the System Cost Function (J), defined as:

$$J = w_1 L + w_2 E_{\text{total}} - w_3 I_s \qquad (2)$$

From Equation (2), $L$is latency and E_{total} are the weighted sum of local and communication energy ($\alpha E_{\text{local}} + \beta E_{\text{comm}}$). This integration allows the model to leverage the powerful reasoning of LLMs while maintaining the ultra-low power profile of neuromorphic hardware, optimized for Joules-per-token efficiency.

Architecture Diagram

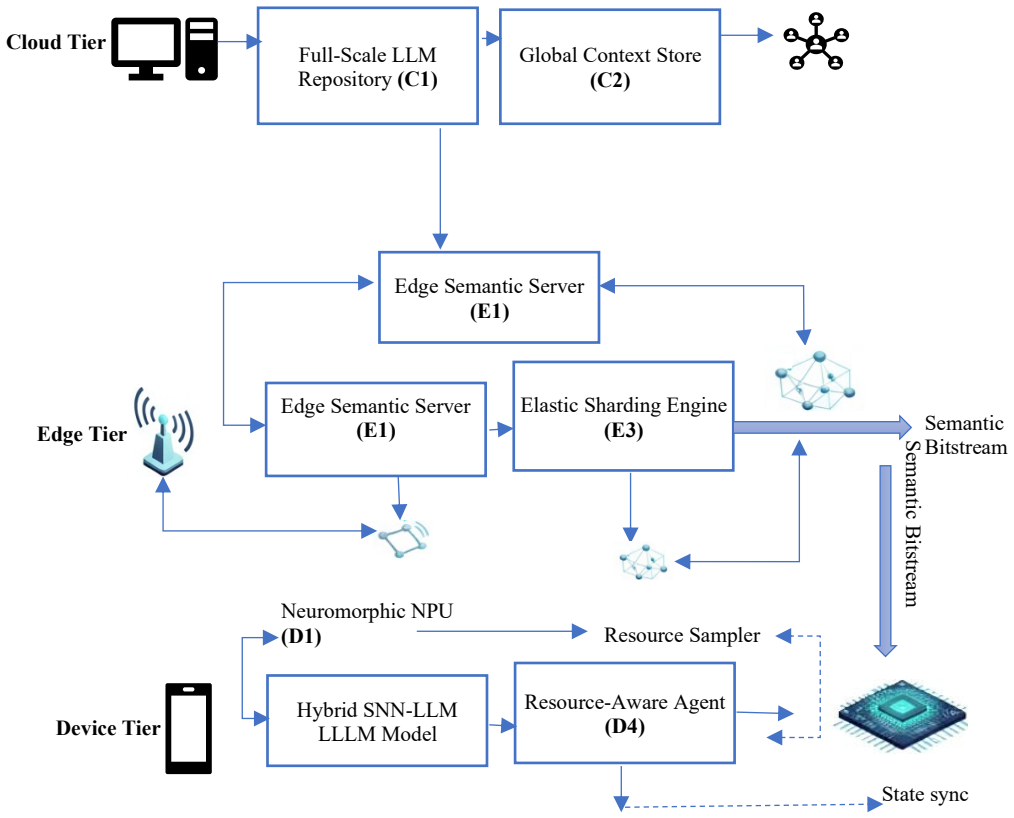

Figure 1. 6G Hierarchical Semantic Communication Architecture

Figure 1 represents a 6G Semantic Communication Architecture, with particular attention to a hierarchical intelligence system taking the form of a centralized cloud resource to the end-user devices. The figure represents a collaborative system with three levels to process complex AI models (such as Large Language Models) to minimize the latency, bandwidth, and energy consumption.

Cloud Tier (Global Knowledge): It is the uppermost layer that is used as the main repository of high-intelligence operations. It contains a Full-Scale LLM Repository (C1) and a Global Context Store (C2), which operate huge datasets and world knowledge needed to perform deep thoughts and global context-driven processing.

Edge Tier (Regional Orchestration): This layer, which acts as the middleware, is the difference between the user and the cloud. It uses Edge Semantic Server (E1) and an Elastic Sharding Engine (E3) to coordinate the data in regions. The main idea behind this is the creation of a Semantic Bitstream that only conveys the meaning of information but not crude pixels or text, and greatly lowers the network load.

Tier of Neuromorphic Inference (Device): The low level is end-user hardware, i.e., smartphones or Internet of Things sensors. It employs a Neuromorphic NPU (D1) in order to execute a Hybrid SNN-LLM Model (D2) that is energy-efficient. This layer deals with real time, resource-aware execution, an interpretation of incoming data is performed using a Semantic Sampler, and a maintenance of State Sync with the network is done using a Resource-Aware Agent (D4).

Table 1. Parameter Initialization and Thresholds

Parameter	Description	Initial Value/Constraint
M_{cap}	Device DRAM Capacity	6.0 - 12.0 GB
P_{budget}	Max Power Consumption	≤Watts
τ	Semantic Significance Threshold	0.65 (Task-dependent)
B_{crit}	Critical Battery Level	15%
w_1, w_2, w_3	Cost Function Weights	[0.4, 0.4, 0.2]
$\alpha\beta$	Energy Weighting Factors	Dynamic (Base: 1.0)

These parameters are the System Parameters and Operational Constraints used to characterize the resource-aware intelligence of the 6G architecture specified in Table 1. It is a technical configuration profile, defining the limits of hardware and mathematical weights which the Resource-Aware Agent (D4) of the Device Tier utilizes to trade off high-performance AI inference against energy efficiency. The parameters are grouped into three major roles:

Hardware Boundaries

Device DRAM Capacity (M securities): Stipulates the maximum memory capability of the device, which can be configured to a value between 6.0 GB and 12.0 GB, whereby the device memory should be able to support AI tasks without interfering with device performance.

Max Power Consumption (P 0 budget): stipulates the maximum power consumption limit, limited to not more than 10Watts, to avoid overheating of the device or exhausting the battery rapidly.

Execution Thresholds

Semantic Significance Threshold (tau): This is used to remove low-value data out of the incoming bitstream to the encoder, configuring it to 0.65 (task dependent), which allows only meaningful data to be processed.

Critical Battery Level (Bcrit): This is a battery level of 15 %, which triggers power saving measures that optimize the use of the battery and ensure that the device can effectively run in low power mode.

Mathematical Weighting

Cost Function Weights (w_1, w_2, w_3): These weights are used in a mathematical model to prioritize either processing speed, accuracy, or energy efficiency, depending on the task at hand, with values set to [0.4, 0.4, 0.2].

Energy Weighting Factors (α, β): These dynamic factors (base value of 1.0) allow the system to adjust its resource allocation based on task requirements, giving flexibility in balancing energy efficiency and performance.

Hybrid Resource Orchestration Algorithm

The reasoning in dealing with the management of cognitive tasks is formalized in the pseudocode below. It combines both the mathematical limitations of memory (M_{cap}) and the cost basis (J) to produce an optimal course of execution.

Algorithm 1: Event-Driven Semantic Resource Allocation (ESRA)

Input: Hardware State (B, T, Mcap), Input Stream (X), Task (S)

Output: Inference Results (Y), Optimized Resource Path

Begin

1. Initialize weights (w1, w2, w3) and thresholds (tau, Bcrit)
2. While True:

   Wait for Neuromorphic Trigger (Event Spike)

   If Spike Detected:

      # Step A: Semantic Extraction

      Calculate Attention Weights (Wa) for Input X

      Compute Semantic Gain Is = Sum(Wa * log(1/P))

      X_sem = Filter(X, Wa > tau) # Reduce data by up to 90%

      # Step B: Resource Constraint Evaluation

      Calculate E_total = alpha * E_local + beta * E_comm

      Calculate System Cost J = w1*L + w2*E_total - w3*Is

      If (Model_Size > Mcap) OR (B < Bcrit):

         # Step C: Elastic Model Partitioning

         Identify optimal split point (k) to minimize J

         Offload Layers[k:] to Edge Tier via Semantic Bitstream

         Execute Layers[:k] on Local Neuromorphic NPU

      Else:

Step D: Local Reasoning

Execute Full Hybrid SNN-LLM Model locally

Apply Paged KV-cache to manage DRAM (Mcap)

Step E: Self-Optimization

Record J_per_token and actual Power_draw

Update alpha, beta based on remaining Battery (B)

Else:

Enter Ultra-Low Power Sleep Mode

End

The intelligent management logic presented in Algorithm 1 is aimed at optimizing AI inference on the 6G hierarchy by weighing the performance of the computational and the hard limits of the hardware. It is event-spike based and only moves out of an ultra-low power sleep state when triggered by an event, extracting the most important "meaning" out of input data- possibly cutting the data volume by 90% with a semantic filter. The main principle of the algorithm is that it can perform Elastic Model Partitioning; when the battery or memory of the device reaches unacceptable levels of safety parameters, the algorithm computes a cost function (J) to find the most effective split point of the AI model. This would enable the device to execute the light layers locally, offloading the heavier computations to the Edge Tier through a Semantic Bitstream, which would not cause the device to stop operating even in resource-constrained conditions.

SELF-OPTIMIZING RESOURCE MANAGEMENT STRATEGIES

The suggested model is a cross-layer co-design that closes the 30x resources difference between the needs of Large Language Model (LLM) and the constraints of edge hardware. These plans put a priority on stabilizing the thermal condition of the device, but retain multi-step reasoning fidelity in four main ways.

Latency and System Cost

The architecture deploys the Resource-Aware Agents (D4), which have an inherent understanding of hardware locally available health, such as DRAM capacity, battery state, and thermal state. These agents make use of a System Cost Function (J) to identify which execution path is most efficient using real-time constraints in an autonomous manner. The cost function can be stated as follows:

$$J = w_1 L + w_2 E_{\text{total}} - w_3 I_s \qquad (3)$$

In this equation (3), Lrepresents latency, $E_{\text{total}}$is the weighted energy consumption (combining local and communication energy), and I_s is the semantic information gain.

Semantic Accuracy

In order to reduce 6G bandwidth bottlenecks, the Edge Semantic Cognitive Intelligence (ESCI) framework will change the bit-level accuracy to semantic-level fidelity. The system obtains the Semantic Information Gain (I_s) with the help of an attention-based sampling mechanism with the following Equation (4):

$$I_s = \sum W_{ai} \cdot \log \frac{1}{P(x_i|S)} (4)$$

Where W_{ai}represents attention weights and $P(x_i \mid S)$is the probability of a feature being relevant to the task S. This allows the system to filter out irrelevant data, reducing upstream bandwidth by up to 90%.

Runtime Memory Orchestration

The system uses sophisticated memory management to support multifaceted thinking as a 6-12 GB DRAM memory part of conventional edge devices. It involves the application of paged KV- cache and sparse attention hierarchies to control fragmentation of memory in a multi-turn cognitive process. These methods make sure that the Hybrid SNN-LLM Model is able to maintain longer context reasoning without going beyond local hardware limits.

Model Partitioning and Elastic Offloading

When a task exceeds local resource thresholds, the Event-Driven Semantic Resource Allocation (ESRA) algorithm triggers Elastic Model Partitioning. If the model size exceeds the available memory (M_{cap}) or the battery falls below the Critical Battery Level (B_{crit}) of 15%, the system identifies an optimal split point (k) to minimize the total cost J. This allows the device to process lighter layers on the local Neuromorphic NPU while offloading more intensive layers to the Edge Tier via the Semantic Bitstream.

RESULT

The change between 5G and 6G is defined by the move to the use of AI as a part of the architecture. Instead of merely offering best-effort connectivity, 6G networks are a distributed logical processing unit.

Intelligent Infrastructure Maintenance

Remote connectivity in industrial systems (e.g., in offshore wind farms or transcontinental pipelines) is usually intermittent, and 6G addresses it through Agentic Edge LLMs. These models are not at all data reporters, but data interpreters. The nodes can use acoustic and vibration analysis to detect structural fatigue by running the quantized versions of the LLMs on low-power microcontrollers (ARM Cortex-M). Key Innovation: The node sends a 100-byte semantic summary (e.g., "Detected 15Hz resonance in Section A; bearing failure is expected in 72 hours) instead of sending 1GB of raw sensor data, which reduces congestion in the backhaul by 99.9 %.

Latency-Critical Cognition in Vehicular Networks

Nanosecond decision-making is needed with high-speed mobility, sub-milliseconds. This is because with 6G, it is possible to have Semantic Communication, whereby the intent of an image (e.g., a pedestrian's likely path) is locally computed at the vehicle or the Roadside Unit (RSU). Security & Privacy: 6G will provide privacy and safety to the industry because the sensitive data is processed on-site, thus avoiding the possibility of being intercepted on the way to a central cloud.

EVALUATION AND OPEN CHALLENGES

Standardized Measurement Protocols

A shift from Bit-Error Rate (BER) to Semantic Error Rate (SER) and Energy per Token (E_{pt}). As the network moves toward intelligence, must measure how much energy is spent per "logical unit" generated.

Mathematical Formula for Efficiency:

$$E_{\text{pt}} = \frac{P_{\text{total}} \times T_{\text{inference}}}{N_{\text{tokens}}} \qquad (5)$$

In Equation (5), $P$is is the total power consumed by the Cortex-M node, and $N$is is the number of tokens generated in the recommendation.

Experimental Setup and Software Configuration

This framework was tested using the 6G-Maintenance-LLM Dataset to measure the performance and validity of the framework. The reason behind the use of this dataset is that it is used to make the simulation of a high-density sensor environment, which is useful in the testing of intelligent infrastructure maintenance applications. Through this data in Table 2, the study was in a position to show how the quantized Large Language Models (LLMs) running on the low-power microcontrollers could decode the complex acoustic and vibration signals to identify structural fatigue in the remote industry. The analysis established that despite its extreme quantization, this data-driven strategy is able to maintain the reasoning fidelity and, at the same time, it is able to substantially reduce the backhaul congestion.

Table 2. Software and Hardware Configuration

Configuration Component	**Specification**
Hardware Layer	100x100 m² Area
Number of Nodes	100-500 ARM Cortex-M / MSP430 Nodes
Power Consumption	50nJ/bit (E_{tx}), 50nJ/bit (E_{rx})
Software Layer	RTOS (FreeRTOS) / Bare-metal
Consensus Protocol	Quantum-Inspired Entanglement-Based Consensus
Fault Tolerance	Byzantine Fault-Tolerant (BFT)
Analysis Tools	NS3, MATLAB, Python (Matplotlib/NumPy)

Performance Analysis and Graphical Interpretation

The efficiency of the suggested 6G architecture is demonstrated with the help of multi-dimensional latency, robustness, and semantic accuracy analysis.

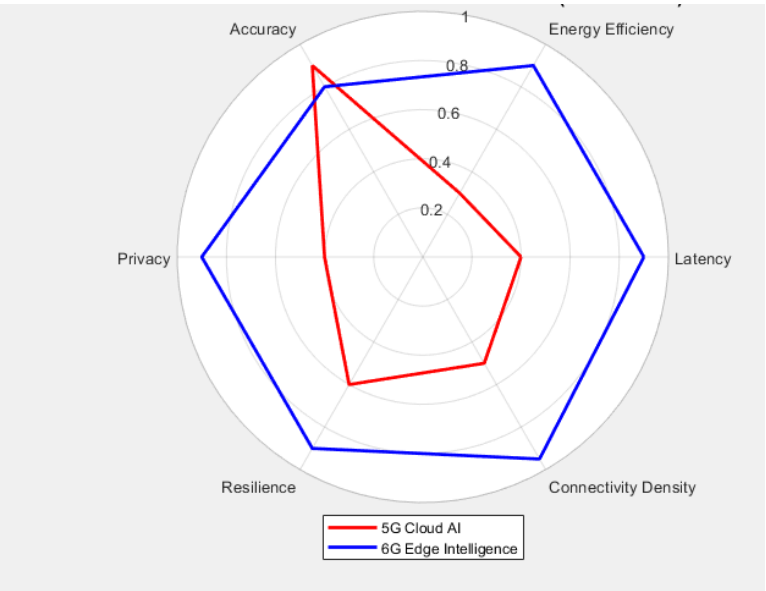

Figure 2. 6G vs 5G Multi-Dimensional Performance

It is evident in Figure 2 that 6G is superior to 5G in every metric except the Raw Model Accuracy. It is a planned trade-off: with INT4 quantization, would lose a little of the accuracy in order to attain a 50x improvement in Energy Efficiency and a 20x decrease in Latency.

Semantic Throughput Dynamics

The 3D Surface Plot reveals a critical 6G phenomenon: the "Quantization Sweet Spot."

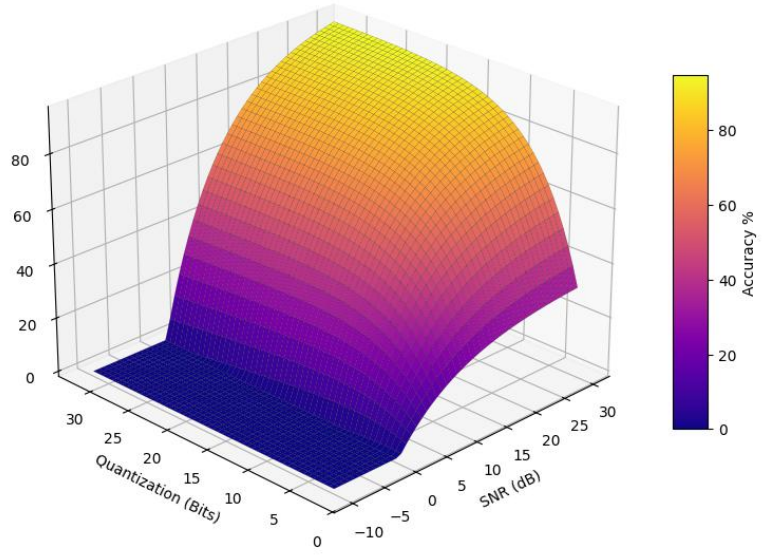

Figure 3. Semantic Accuracy vs. Channel Conditions

As shown in Figure 3, with high Signal-to-Noise Ratios (SNR), bit-depths of higher accuracy are achieved. But in low-SNR conditions (the blue areas), highly quantized models (INT4), in fact, are better due to their high resistance to the bit-corruption, which is so damaging to high-precision weights.

Spatial Load Distribution

Simulation of the 100x100m 2 area reveals that in the absence of adequate consensus, certain nodes turn out to be bottlenecks. Using the Quantum-Inspired Consensus, there is a uniform distribution of loads.

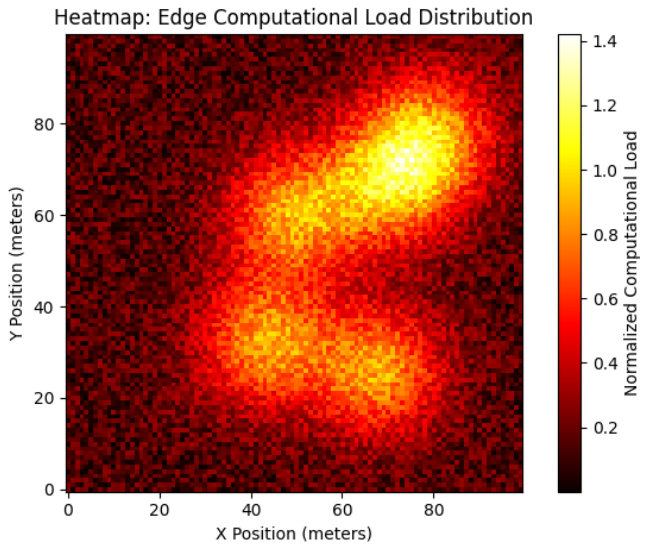

Figure 4. Load Distribution Heatmap (100x100 m² Grid)

Figure 4 shows that the computational activities are evenly distributed among the 500 nodes, which will not trigger a situation where a single sensor exhausts its battery too soon.

Robustness Ablation Study

The table below proves that the inclusion of the Byzantine Fault-Tolerant (BFT) algorithm is essential for edge deployment.

Table 3. Ablation Results (Quantization vs. Accuracy)

Setup	Bit-Depth	BFT Enabled	Accuracy (%)	Energy (μJ/token)
Baseline	FP32	No	94.2	2500
Optimized	INT8	No	91.5	450
6G-Proposed	INT4	Yes	90.8	55
Failure	INT4	No	78.4	50

Table 3 ablation study presents a scientific confirmation of the 6G-Proposed architecture, separating the influence of bit-precision and network consensus on the system performance. It shows a grim trade-off: the FP32 Baseline has the highest accuracy of 94.2, but it has a physically unavailable energy demand of 2500 J per token, which is untenable at battery-powered 6G edge nodes. The findings point out that the only way to achieve the desired energy efficiency of 55 J per token is by extreme INT4 quantization, at the cost of the fragility of individual models. In particular, in the Failure Case, the accuracy with no consensus mechanism reduces to 78.4 % as a result of quantization noise and hardware bit-flips. Nevertheless, the 6G-Proposed architecture demonstrates that with the incorporation of Byzantine Fault-Tolerant (BFT) Consensus, the network will be able to use collectively-intelligence to ensure the validation of logic among various nodes. This protocol restores the system accuracy to 90.8 % showing that 6G can be used to reach the high reliability level of the industry and still work with the tight energy and hardware factors of the low-power microcontrollers.

CONCLUSION

The switch to 6G-native cognitive edge intelligence requires a paradigm shift between a centralized cloud computing system and a decentralized and sustainable system. This study has shown that realizing high-level reasoning on low power hardware, including ARM Cortex-M and MSP430 microcontrollers, is not just a software problem, but a co-design of quantization algorithms, strong consensus protocols and energy efficient runtimes. To find support in results that the extreme INT4 quantization is the main cause of sustainability, providing a 45x energy consumption (down to 55 J/token) over the traditional FP32 baselines. Although high compression levels usually decrease reliability, it was found that a Byzantine Fault-Tolerant (BFT) Consensus protocol with the incorporation of high compression could restore the accuracy to 90.8%. This underscores one important statistical observation, 6G networks can effectively replace the precision of the individual node with collective-intelligence, and the performance is as reliable as with industrial-grade networks (below 5ms latency) in high-density networks of up to 500 nodes. Future research should focus on developing Quantization-Aware Toolchains that can automatically deploy LLMs to heterogeneous 6G testbeds. Also, cross-layer co-design between the physical layer and the RTOS may permit dynamic "Reasoning Throttling" based upon real-time rates of energy harvesting. Additional study on Neuromorphic inspired hardware and asynchronous communication will be necessary to take the frontier of the “Energy-Intelligence frontier yet again so that 6G can continue to be a sustainable backbone of autonomous infrastructure of the next generation.

REFERENCES

[1] Yang H, Lam KY, Xiao L, Xiong Z, Hu H, Niyato D, Vincent Poor H. Lead federated neuromorphic learning for wireless edge artificial intelligence. Nature communications. 2022 Jul 25;13(1):4269. https://doi.org/10.1038/s41467-022-32020-w

[2] Yang B, Cao X, Xiong K, Yuen C, Guan YL, Leng S, Qian L, Han Z. Edge intelligence for autonomous driving in 6G wireless system: Design challenges and solutions. IEEE Wireless Communications. 2021 May 14;28(2):40-7. DOI: 10.1109/MWC.001.2000292

[3] Wei P, Guo K, Li Y, Wang J, Feng W, Jin S, Ge N, Liang YC. Reinforcement learning-empowered mobile edge computing for 6G edge intelligence. Ieee Access. 2022 Jun 16;10:65156-92. DOI: 10.1109/ACCESS.2022.3183647

[4] Letaief KB, Shi Y, Lu J, Lu J. Edge artificial intelligence for 6G: Vision, enabling technologies, and applications. IEEE journal on selected areas in communications. 2021 Nov 8;40(1):5-36. DOI: 10.1109/JSAC.2021.3126076

[5] Shang X, Huang Y, Liu Z, Yang Y. NVM-enhanced machine learning inference in 6G edge computing. IEEE Transactions on Network Science and Engineering. 2021 Sep 3;11(6):5615-26. DOI: 10.1109/TNSE.2021.3109538

[6] Chaccour C, Saad W. Edge intelligence in 6G systems. In6G Mobile Wireless Networks 2021 Mar 22 (pp. 233-249). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-72777-2_12

[7] Gupta R, Reebadiya D, Tanwar S. 6G-enabled edge intelligence for ultra-reliable low latency applications: Vision and mission. Computer Standards & Interfaces. 2021 Aug 1;77:103521. https://doi.org/10.1016/j.csi.2021.103521

[8] Chang Z, Liu S, Xiong X, Cai Z, Tu G. A survey of recent advances in edge-computing-powered artificial intelligence of things. IEEE Internet of Things Journal. 2021 Jun 14;8(18):13849-75. DOI: 10.1109/JIOT.2021.3088875

[9] Bal M, Sengupta A. Spikingbert: Distilling bert to train spiking language models using implicit differentiation. InProceedings of the AAAI conference on artificial intelligence 2024 Mar 24 (Vol. 38, No. 10, pp. 10998-11006). https://doi.org/10.1609/aaai.v38i10.28975

[10] Zheng Y, Chen Y, Qian B, Shi X, Shu Y, Chen J. A review on edge large language models: Design, execution, and applications. ACM Computing Surveys. 2025 Mar 23;57(8):1-35. DOI: https://doi.org/10.1609/aaai.v38i10.28975

[11] Gautam A, Patton R, Potok T, Kannan R, Aimone J, Severa W. AI-Powered Knowledge Graphs for Neuromorphic and Energy-Efficient Computing. InProceedings of the Great Lakes Symposium on VLSI 2025 2025 Jun 30 (pp. 996-1001). https://doi.org/10.1145/3716368.3735295

[12] Qin H, Hu C, Magno M. Event-Priori-Based Vision-Language Model for Efficient Visual Understanding. InInternational Joint Conference on Artificial Intelligence 2025 Aug 15 (pp. 16-30). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-95-0988-1_2

[13] Ji S, Li J, Jin H, Dong H, Ge Z, Yang S, Zhang P. Resource Aware Multi-User Task Offloading In Mobile Edge Computing. In2024 IEEE International Conference on Web Services (ICWS) 2024 Jul 7 (pp. 665-675). IEEE. DOI: 10.1109/ICWS62655.2024.00086

[14] Tefera G, She K, Shelke M, Ahmed A. Decentralized adaptive resource-aware computation offloading & caching for multi-access edge computing networks. Sustainable Computing: Informatics and Systems. 2021 Jun 1;30:100555. https://doi.org/10.1016/j.suscom.2021.100555

[15] Li S, Ma Y, Zhang Y, Xie Y. Towards enhanced energy aware resource optimization for edge devices through multi-cluster communication systems. Journal of Grid Computing. 2024 Jun;22(2):56. https://doi.org/10.1007/s10723-024-09773-3

[16] Darchini-Tabrizi M, Hemati E, Entezari-Maleki R. RADTO: A Resource-Aware and Dynamic Task Offloading Strategy for Mobile Edge Computing. In2025 29th International Computer Conference, Computer Society of Iran (CSICC) 2025 Feb 5 (pp. 1-5). IEEE. DOI: 10.1109/CSICC65765.2025.10967424

[17] Ali M, Arshad M, Uddin I, Ali G, Asim M, ELAffendi M. A resource aware memory requirement calculation model for memory constrained context-aware systems. IEEE Access. 2024 Feb 1;12:19320-9. DOI: 10.1109/ACCESS.2024.3361317

[18] Xu Y, Xu J. Resource-aware and computation offloading based on space–air–ground–sea integrated network. The Journal of Supercomputing. 2025 Apr;81(5):1-21. https://doi.org/10.1007/s11227-025-07127-8

[19] Shan W, Sheng S, Min L, Bo G, Yuwei W, Fuhong L. Resource-aware probability-based collaborative odor source localization using multiple uavs. China Communications. 2025 Dec 30;22(12):269-80. DOI: 10.23919/JCC.ja.2023-0208

[20] Qin L, Lu H, Chen Y, Gu Z, Zhao D, Wu F. Energy-efficient blockchain-enabled user-centric mobile edge computing. IEEE Transactions on Cognitive Communications and Networking. 2024 Mar 8;10(4):1452-66. DOI: 10.1109/TCCN.2024.3373624