ACCEPTED FROM OPEN CALL

ENWAR: A RAG-Empowered Multi-Modal LLM Framework for Wireless Environment Perception

Ahmad M. Nazar, Abdulkadir Celik, Mohamed Y. Selim, Asmaa Abdallah, Daji Qiao, and Ahmed M. Eltawil

Abstract

Large language models (LLMs) hold significant promise in advancing network management and orchestration in sixth-generation (6G) and beyond networks. However, existing LLMs are limited in domain-specific knowledge and their ability to handle multi-modal sensory data, which is critical for real-time situational awareness in dynamic wireless environments. This article addresses this gap by introducing ENWAR,[1] an ENvironment-aWARe retrieval-augmented generation (RAG)-empowered multi-modal LLM framework. ENWAR seamlessly integrates multi-modal sensory inputs to perceive, interpret, and cognitively process complex wireless environments to provide human-interpretable situational awareness. ENWAR is evaluated on the global positioning system (GPS), light detection and ranging (LiDAR) sensors, and camera modality combinations of the DeepSense6G dataset with state-of-the-art LLMs such as Mistral-7b/8x7b and LLaMa3.1-8/70/405b. Compared to general and often superficial environmental descriptions of these vanilla LLMs, ENWAR delivers richer spatial analysis, accurately identifies positions, analyzes obstacles, and assesses line-of-sight (LoS) between vehicles. Results show that ENWAR achieves key performance indicators of up to 70% relevancy, 55% context recall, 80% correctness, and 86% faithfulness, demonstrating its efficacy in multi-modal perception and interpretation.

Introduction

Generative artificial intelligence (AI), with its ability to synthesize, adapt, and contextualize data, is poised to drive the evolution of sixth-generation (6G) and beyond networks [1]. Among generative models, large language models (LLMs) have emerged as the most transformative, redefining how machines comprehend and generate human language. Built on transformer architectures and powered by attention mechanisms, LLMs leverage large-scale pretraining to excel in natural language understanding, reasoning, and decision support [2]. Their adaptability and scalability make them ideal for dynamic, complex systems, positioning LLMs as key enablers of AI-native wireless intelligence and self-optimizing 6G networks, ultimately paving the way toward zero-touch network and service management (ZSM).

However, the technical demands of next-generation networks differ greatly from legacy generations. Future networks are expected to operate with massive antenna arrays at significantly higher frequencies, wherein wireless channels become less probabilistic and more deterministic and exhibits geometric propagation characteristics. This shift introduces daunting mobility challenges such as tracking narrow beams, blockage mitigation through timely handover management, and seamless service migration. Sensing functionalities and environmental awareness are essential for ZSM to effectively navigate this new terrain.

In this context, multi-modal integrated sensing and communication (ISAC) represent coherent fusion of disparate data streams from various sensors (e.g., light detection and ranging (LiDAR), radars, cameras, global position system (GPS), etc.), unlocking critical capabilities such as environment mapping, object/human detection and classification, urban planning, localization, and tracking. These sensing functionalities collectively lay the foundations of digital twins (DTs); a dynamic and near real-time virtual replica of 6G networks, providing contextual and site-specific insights into the spatio-temporal characteristics of wireless environments [3]. DTs are crucial in optimizing network performance, enabling real-time decision-making, and enhancing overall situational awareness, making it an integral component of the future telecom ecosystem.

Nonetheless, LLMs mainly operate in text-based modalities, limiting their ability to process multi-modal sensory inputs — an essential requirement for situation-aware networks where real-world comprehension extends beyond text. Moreover, their vast but generic pretraining often falls short on domain-specific tasks and can suffer from outdated knowledge or hallucinations due to their reliance on probabilistic pattern matching rather than true reasoning [4]. To mitigate these issues, retrieval-augmented generation

[1] Enwar is a common name in Turkic and Arabic cultures, meaning more enlightened, insightful, and intellectual; herein referring to a multi-modal LLM providing deep situational and contextual insights into the environment.

ISSN: 0163-6804
Digital Object Identifier: 10.1109/MCOM.001.2500261

Ahmad M. Nazar, Mohamed Y. Selim, and Daji Qiao are with Iowa State University (ISU), USA; Abdulkadir Celik is with the University of Southampton, UK.; Asmaa Abdallah and Ahmed M. Eltawil are with the King Abdullah University of Science and Technology (KAUST), Saudi Arabia

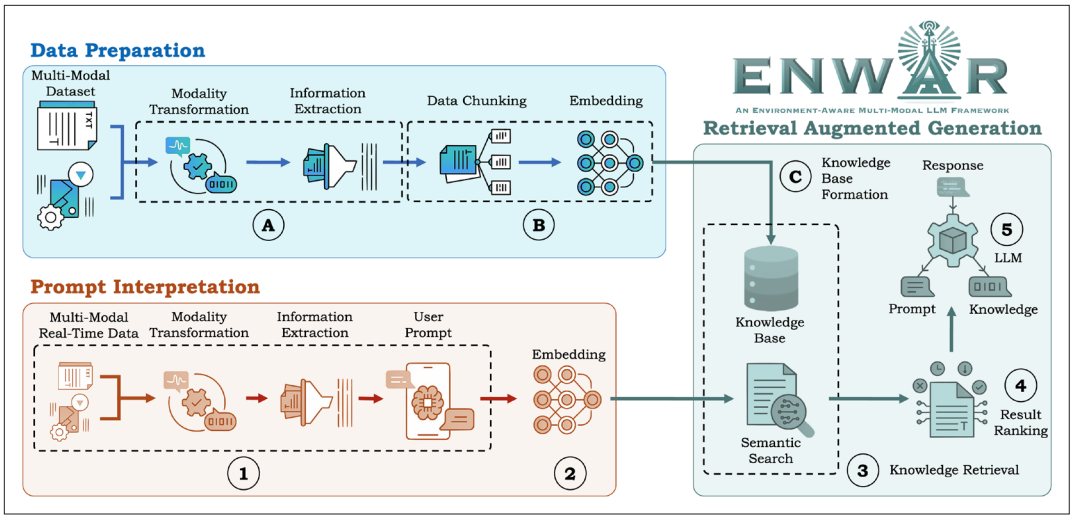

FIGURE 1. ENWAR workflows: multi-modal RAG formation (Steps A-C); and prompt interpretation, knowledge retrieval, and response generation (Steps 1-5).

(RAG) frameworks integrate external knowledge retrieval into the generative process, enabling LLMs to access domain-specific or real-time data sources. By retrieving semantically relevant document chunks, RAG systems enhance factual accuracy and contextual grounding, bridging the gap between general-purpose LLMs and specialized 6G network needs. However, while RAG improves domain expertise, it alone cannot address the needs of ISAC multi-modal for fully situational-aware 6G networks.

The integration of LLMs into wireless systems has been preliminarily explored in [5], which focused on textual data and lacked RAG-based reasoning, limiting applications to basic telecom chatbots. The work in [6] demonstrates an environment-aware vehicular assistant, showcasing the potential of reasoning-based perception within wireless contexts. Vision–language models for autonomous driving orchestration are further examined in [7]. WirelessLLM [8] incorporates domain knowledge for spectrum sensing and protocol understanding. In contrast, Xu *et al.* [9] propose modular edge LLMs with perception, grounding, and alignment stages for 6G tasks. Similarly, NetOrchLLM [10] leverages large models for orchestrating dense networks and dynamic environments. Although recent studies [11, 12] discuss the promise of multi-modal LLMs, these works lack concrete case studies or real-world demonstrations of multi-modal grounding, leaving open the need for frameworks like ENWAR[2] that operationalize these capabilities in wireless settings.

We address this gap in the wireless literature by introducing ENWAR, an Environment-aware RAG-empowered multi-modal large language model (MLLM) framework that leverages multi-modal sensory data to perceive, interpret, and cognitively process complex wireless environments. ENWAR's human-interpretable situational awareness is crucial for both sensing and communication applications, where real-time environmental perception can significantly enhance system performance and reliability. Unlike prior vision–language or wireless perception models that rely solely on end-to-end feature fusion, ENWAR uniquely integrates retrieval-augmented generation to ground multi-modal sensory inputs, such as LiDAR, camera, and GPS data, within a structured wireless knowledge base. This enables the model to reason over historical contexts, infer spatial relationships, and produce domain-consistent responses that reflect wireless propagation and blockage dynamics. By coupling RAG-driven memory retrieval with multi-modal grounding, ENWAR transforms otherwise generic LLMs into context-aware reasoning engines for 6G environment perception.

The following sections outline ENWAR's workflow and introduce four key performance indicators (KPIs): answer relevancy, context recall, correctness, and faithfulness. ENWAR's performance is evaluated using Mistral-7b/8x7b and LLaMA3.1-8/70/405b models across GPS, LiDAR, and camera modalities in vehicle-to-vehicle scenarios from the DeepSense6G dataset [13]. While off-the-shelf pre-trained LLMs (a.k.a, Vanilla LLMs) provide only generic environment descriptions, ENWAR offers contextually grounded spatial reasoning, accurately identifying entities, distances, and potential obstacles, and assessing line-of-sight (LoS) between communicating vehicles. Quantitatively, ENWAR achieves up to 70% relevancy, 55% context recall, 80% correctness, and 86% faithfulness, demonstrating its strength in multi-modal understanding. The article concludes with future directions on extending these perception and cognition capabilities toward intelligent, 6G-aware wireless systems.

Unlike prior vision–language or wireless perception models that rely solely on end-to-end feature fusion, ENWAR uniquely integrates retrieval-augmented generation to ground multi-modal sensory inputs, such as LiDAR, camera, and GPS data, within a structured wireless knowledge base.

An Overview of ENWAR Framework

As illustrated in Fig. 1, ENWAR is comprised of two primary workflow pipelines:

1. Multi-modal RAG formation (Steps A-C), and
2. Prompt interpretation, knowledge retrieval, and response generation (Steps 1–5); which are described in the following sections along with KPIs.

Multi-Modal RAG Formation

Ⓐ **Dataset Preprocessing and Modality Transformation:** ENWAR is designed to seamlessly accommodate diverse sensor modalities by preprocessing and transforming them into a unified textual format that can be effectively processed by LLMs. For instance, GPS data undergoes transformation from raw spatial coordinates into

[2] ENWAR's code is available at https://github.com/anazar99/Enwar1.0

By consolidating various sensory data into a textual format (e.g, JavaScript object notation (JSON) format), ENWAR ensures that LLMs can cohesively process and interpret multi-modal inputs, enhancing the model's ability to generate contextually aware and reliable outputs.

textual descriptions that provide insights such as relative distances, directional bearings, and movement patterns, offering a richer contextual understanding of spatial relationships.

Visual data is processed through an image-to-text conversion model that extracts key visual elements and translates them into LLM interpretable natural language descriptions. The use of instructional prompts ensure that the generated textual outputs are contextually relevant and sufficiently detailed to accurately represent the visual information.

Point cloud data from LiDARs, another complex modality, is processed by feature extraction models (e.g., ResNet) to identify salient environmental elements. Object detection and classification systems are then employed to recognize key entities (e.g., pedestrians, vehicles), which are subsequently converted into textual descriptions.

The final step in the preprocessing pipeline involves synthesizing the transformed data from all modalities into a unified textual representation. By consolidating various sensory data into a textual format (e.g, JavaScript object notation (JSON) format), ENWAR ensures that LLMs can cohesively process and interpret multi-modal inputs, enhancing the model's ability to generate contextually aware and reliable outputs. This synthesis is pivotal for enabling the framework to make informed decisions in complex environments.

Ⓑ **Text Chunking and Embedding:** ENWAR's next critical step is to segment the sensory data into manageable chunks and convert these chunks into numerical embeddings. In this way, LLMs can efficiently process and interpret textual information, especially when handling large datasets from diverse sensor modalities.

Chunking involves breaking down the preprocessed text into smaller, equal-sized and contextually coherent units. This is essential as LLMs have token limits, meaning that excessively large text inputs cannot be processed effectively. Segmentation ensures that the model can focus on relevant parts of the data without losing contextual integrity. For instance, GPS data may be chunked based on time intervals or location changes, while visual and point cloud descriptions could be divided based on objects detected or spatial regions.

Once the data is chunked, it is passed to a General Text Embeddings (GTE) model to convert each chunk into a dense vectorized format; a numerical representation of the text capturing its semantic content. These embeddings serve as a structured and machine-readable format that encodes the underlying text's meaning. Vectorization enables LLMs to tokenize and process the data, establishing relationships between different chunks based on their semantic similarity.

The alignment across modalities is achieved during tokenization and embedding, where padding ensures equal-length tokenized chunks and consistent semantic representation. This uniformity enables effective handling of redundancy, conflicts, and synergies across modalities through the LLM's interpretative and generative capabilities. By analyzing the transformed and integrated data, the LLM reconciles discrepancies, identifies shared patterns, and leverages each modality's strengths. This cohesive approach ensures robust, conflict-free environment sensing and a unified understanding of the environment.

Ⓒ **Domain-Specific Knowledge Base Generation:** ENWAR's ability to deliver precise and context-aware responses is largely dependent on its robust domain-specific knowledge base. The RAG knowledge base is built using embeddings from various sensor modalities, which are stored and indexed with the Facebook AI similarity search (FAISS) (https://ai.meta.com/tools/faiss/) library. During construction, FAISS clusters embeddings semantically and indexes hierarchically for high precision and low latency. Moreover, consistent chunking and padding ensure seamless multi-modal data integration, supporting efficient, context-aware prompt handling. The knowledge base's modular design, organized by sensor modality and contextual relevance, enhances scalability and retrieval efficiency by narrowing searches to specific subsets. All these aspects of RAG formation enables real-time decision-making through swift retrieval of the relevant data and ensures that ENWAR remains adaptable and responsive to a wide range of scenarios, enhancing its performance in dynamic and complex wireless environments.

PROMPT INTERPRETATION AND RESPONSE GENERATION

① **Prompt Preprocessing and Modality Transformation:** This step closely mirrors the procedures in Step-Ⓐ: the user prompt is preprocessed by transforming its components and any real-time multi-modal sensory data into a unified, standardized format suitable for LLMs. This ensures the prompt is properly aligned with the knowledge base, allowing for seamless interaction with the model's retrieval mechanisms.

② **Prompt Text Embedding:** Similarly, this step follows the procedures in Step-\textcircled{\footnotesize B}: the preprocessed prompt is converted into numerical embeddings, ensuring that it can be efficiently processed and compared to the vectorized data in the knowledge base. This transformation facilitates accurate retrieval of relevant information, streamlining the prompt's interaction with the model's generative components.

③ **Semantic Search and Knowledge Retrieval:** Once the user prompt has been transformed into embeddings, ENWAR performs semantic search to retrieve the most relevant information from its domain-specific knowledge base in the RAG framework. This process identifies entries that closely match the prompt by calculating the semantic similarity between the prompt and the embedded data in the knowledge base. Through the knowledge base, FAISS optimizes vector similarity searches by implementing hierarchical indexing and clustering, allowing ENWAR to handle large-scale embeddings efficiently. This ensures that top-ranked result retrieval operations remain low-latency even as data volume grows. As detailed next, the top-ranked results, which are contextually aligned with the prompt, are then selected for further processing.

④ **Result Ranking:** ENWAR ensures relevance by ranking results according to their section headers, prioritizing the most contextually appropriate portions of the knowledge base. This refined search mechanism optimizes retrieval by focusing on the most pertinent content. Since some contexts may have similar vectorized embeddings, ENWAR concentrates on the top-p percentile to

effectively filter out less relevant data, with p = 95 used throughout the system to anchor the retrieval process in the highest-ranking results. This approach enhances both the precision and relevance of the retrieved information for more accurate and contextually appropriate responses.

⑤ **Response Generation:** Once the top-ranked results from the semantic search are identified, they provide the LLM essential context to generate a coherent and contextually appropriate response. These results serve as the foundation upon which the LLM builds its output, ensuring that the generated response is accurate and relevant to the user's prompt.

The LLM processes the vectorized embedding of the user prompt along with the retrieved context from the knowledge base. It integrates information from multiple sources, such as GPS coordinates, LiDAR data, and visual descriptions, to construct a detailed representation of the environment. This may involve detecting vehicles, their locations, and describing physical aspects of the surroundings in relation to the prompt. To further enhance the generation process, ENWAR employs top-p sampling to strike a balance between accuracy and diversity in responses, effectively filtering out irrelevant outputs while maintaining contextual richness.

Beyond basic description, the LLM infers interactions among environmental elements. Using GPS data and contextual cues, it can anticipate how vehicles and surroundings influence one another. The model aligns its responses with user-defined tasks, delivering comprehensive situational awareness by identifying key entities, their positions, and potential interactions. By synthesizing and grounding multi-modal data within the retrieved context, the LLM produces detailed, actionable insights that support informed decision-making in dynamic, complex environments.

KEY PERFORMANCE INDICATORS

Evaluating the performance of ENWAR requires assessing its output based on both general benchmarks and domain-specific metrics. Standard benchmarks such as the General Language Understanding Evaluation (GLUE) and Massive Multitask Language Understanding (MMLU offer a broad assessment of an LLM's capabilities across various metrics such as answer relevancy, factual correctness, and hallucinations avoidance. However, RAG-based systems require a more targeted evaluation due to their reliance on domain-specific contexts, and in our case, specifically tailored multi-modal data. Following the RAGAS (https://docs.ragas.io) framework, we ensure a comprehensive and accurate evaluation of ENWAR's performance through following KPIs:

- ***Answer Relevancy*** (AR) measures how well ENWAR's responses align semantically with the user's prompt to ensure the ability of generating contextually relevant responses for actionable insights. Denoting E_{pi} and E_{ti} as the embedding vectors of ith generated prompt and the relevant ground truth of the data sample, respectively; cosine/semantic similarity, $-1 \leq \cos(\mathbf{E}_{pi}, \mathbf{E}_{ti}) \leq 1$, measures how semantically similar two texts are based on their vector representations, i.e., 1: perfectly similar (aligned), 0: no similarity, −1: completely dissimilar (opposite). Accordingly, AR is evaluated by calculating the average cosine similarity as follows

$$\text{AR} = \frac{1}{N}\sum_{i=1}^{N}\cos(\mathbf{E}_{p_i}, \mathbf{E}_{t_i}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{E}_{p_i}\cdot\mathbf{E}_{t_i}}{\|\mathbf{E}_{p_i}\|_2\,\|\mathbf{E}_{t_i}\|_2}. \quad (1)$$

- ***Context Recall*** evaluates whether ENWAR correctly recalls information from the knowledge base that is relevant to the prompt and verifies how much of the response can be attributed to the correct context. High context recall is critical in real-world scenarios where situational-awareness depends on the accurate retrieval of modality and domain-specific data, such as identifying potential obstacles in a dynamic environment. It is calculated by normalizing the alignment extent of the retrieved contexts within the ground truth with the number of sentences in the ground truth.
- ***Correctness Score*** combines semantic similarity and factual accuracy to measure how well ENWAR interprets and integrates multi-modal data. Correctness directly impacts the system's reliability in applications, wherein even minor inaccuracies could result in suboptimal decisions. Denoting the embedding vector of ith generated answer by $\mathbf{E}_{a_i}$ and F_1 score as a metric of factual correctness, the overall correctness score is given by

$$\text{Correctness} = \omega\cos(\mathbf{E}_{a_i}, \mathbf{E}_{t_i}) + (1-\omega)F_1, \quad (2)$$

where weighting parameter $0 \leq \omega \leq 1$ — RAGAS sets $\omega = 0.25$ as default — ensures that response assessments are both factually accurate and contextually appropriate.
- ***Faithfulness*** evaluates the consistency of the generated answers with the retrieved context. A response is considered faithful if all claims align with the retrieved data, ensuring the output does not contain unsupported or fabricated information. This metric is crucial in environmental sensing applications to ensure that responses are fully attributed to the appropriate environment and contains minimal hallucinations. Faithfulness checks whether the claims made in the output can be logically deduced from the given context and is given by

$$\text{Faithfulness} = \frac{|N_{G_c}|}{|N_C|}, \quad (3)$$

where N_{G_c} is the number of claims in the generated answer that can be inferred from the given context and N_C is the total number of claims in the generated answer.

ENWAR SETUP BREAKDOWN AND A CASE STUDY

This section provides a detailed breakdown of the ENWAR setup and offers a qualitative performance comparison with vanilla LLMs, highlighting the advantages of integrating multi-modal data and knowledge retrieval within ENWAR.

A BREAKDOWN OF ENWAR SETUP

DeepSense6G Dataset: To evaluate ENWAR's performance, we utilize a large-scale, real-world multi-modal sensing and communication dataset

Evaluating the performance of ENWAR requires assessing its output based on both general benchmarks and domain-specific metrics.

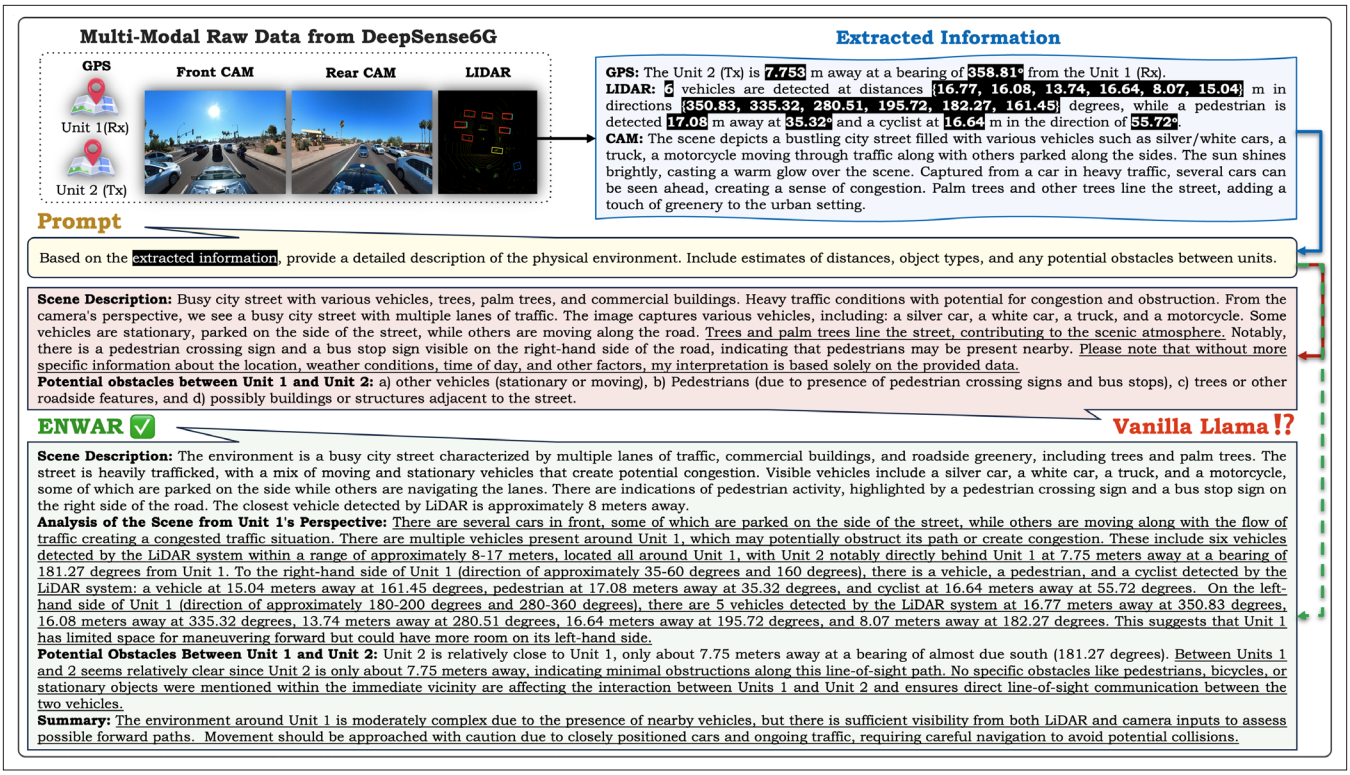

FIGURE 2. Illustration of the case study scene with raw data, extracted information, generated prompt, and responses from Vanilla Llama and ENWAR.

[13]. DeepSense6G provides a robust platform for testing ENWAR's ability to interpret complex spatial and environmental data. For our evaluation, we focus on Scenario 36, which includes GPS coordinates of a vehicle equipped with four signal receivers, its captured 360-degree LiDAR point clouds and front–rear camera frames, and the GPS coordinates of another vehicle equipped with a transmitter sampled every 100ms, yielding thousands of temporally dense frames representative of real-time vehicular motion. We meticulously identified 180 diverse scenes — 30 reserved for testing — covering urban settings with varying numbers of pedestrians, cyclists, and vehicles; an exemplary scene is shown in Fig. 2. These scenes collectively span highways, parking lots, intersections, empty streets, and main roads, ensuring broad environmental diversity and statistical representativeness of dynamic 6G scenarios. Scene annotations were performed using open-source labeling software and validated through cross-checking the outputs of the super fast accurate 3D SFA3D (https://github.com/maudzung/Super-Fast-Accurate-3D-Object-Detection) and InstructBLIP models, to ensure cross-modality alignment. This semi-automated pipeline establishes reproducibility while remaining scalable to real-time environments through integrated perception modules.

Modality Transformation and Information Extraction: For all selected scenes, ENWAR extracts latitude and longitude coordinates from GPS inputs to determine the positions and relative bearings of two vehicles, which are then converted into textual format for seamless integration into ENWAR's prompt. For front-rear images, ENWAR performs image-to-text transformation to generate a textual description of the visual content by using InstructBLIP trained on Vicuna-7b and optimized for visual-tuned instructions [14].

For LiDAR point clouds, ENWAR leverages SFA3D model for object detection and analysis. SFA3D was modified to extract object information, including locations and bearings relative to the sensor, and converts this data into text to describe the environment. SFA3D utilizes a ResNet-based keypoint feature pyramid network (KFPN) for reliable LiDAR object detection, transforming 3D point clouds into Birds-Eye-View images, which are then processed to identify objects with high confidence, providing detailed information such as positions, dimensions, and orientations.

The extracted information from each modality is hard-coded into a template [c.f., white text highlighted with black background in Fig. 2 to be utilized during the prompting and grounding process. Since object types and potential blockages are not readily available labels within the DeepSense6G dataset, we manually create ground truth text for all 180 scenes by correcting extracted information if necessary and/or adding missing details.

Instructional Text Prompt: The prompt includes *extracted information* of each scene and specifies predefined tasks, guiding ENWAR to accurately analyze the wireless environment, detect potential blockages, and generate relevant insights based on the processed multi-modal inputs [yellow box in Fig. 2.

Chunking and Embedding: ENWAR employs the gte-large-en-v1.5 embedding model from AliBaba-NLP for knowledge base construction [15]. The model supports a context length of 8,192 tokens. Textual data is divided into 1,024-character chunks with a 100-character overlap to preserve continuity, then tokenized and padded to equal length for consistent embedding generation.

Knowledge Base Creation and Performance Evaluation: ENWAR utilizes FAISS to cre-

ate knowledge bases — one for each modality combination — and its efficient search/retrieval through the top-95% ranking explained above. ENWAR's performance was thoroughly evaluated by running RAGAS framework across LLMs such as Mistral-7b/8x7b (https://mistral.ai) and Llama3.1-8/70/405b (https://www.llama.com), with model sizes ranging from 7b to 405b parameters. We utilized these pre-trained LLMs hosted on Hugging Face servers, enabling efficient inference for near-real-time performance. For comparison, baseline versions of these LLMs were used to benchmark the performance of vanilla LLMs against ENWAR, particularly in generating detailed and accurate responses. To enhance multi-modal perception tasks by balancing interpretive ability, creativity, conciseness, and response relevance; the evaluation phase used the following key hyperparameter values for all LLMs: `max_new_tokens` = 4096, `temperature` = 0.5, and `repetition_penalty` = 1.15.

A Comparative Analysis of an Inference Case Study

In this section, we evaluate the perception capabilities of Vanilla Llama and ENWAR in processing, analyzing, and interpreting spatial relationships between objects, as well as inferring potential obstacles between the two units. As shown in Fig. 2, both models were tasked with generating a detailed description of a busy city street scene. To provide more descriptive insights, we selected a specific scenario featuring congested traffic, including cars, motorcycles, pedestrians, and other stationary objects along the road.

While Vanilla LLaMA provides a general scene description and identifies visible entities, its response remains superficial; offering basic distance and direction details without analyzing spatial relationships or inferring how obstacles impact movement or communication between the units.

By contrast, as shown in Fig. 2, ENWAR delivers a detailed, contextually grounded breakdown of spatial dynamics from Unit 1's perspective. It accurately identifies vehicles, pedestrians, and cyclists, analyzes potential obstructions, and suggests maneuvering strategies within congested environments. Crucially, it assesses LoS communication and identifies signal blockages, demonstrating advanced perception capabilities that set it apart from the generic output of Vanilla LLaMA.

At this stage, it is crucial to compare the corresponding KPIs for the scene depicted in Fig. 2. As summarized in Table 1, where context recall is omitted as a RAG-specific metric, ENWAR consistently outperforms Vanilla LLaMA across relevancy, correctness, and faithfulness, confirming its superior contextual reasoning and reliability. Its single-modality inference times are 100ms for GPS, 100ms for LiDAR, and 2.5s for camera inputs. Although image-to-text translation dominates latency, scene elements such as traffic, weather, and landscape typically vary on the order of seconds, meaning visual reprocessing is not always required. LiDAR and GPS updates, by contrast, provide millisecond-level quantitative feedback, enabling ENWAR to use windowed tracking for efficient real-time updates. These inference times can be further optimized through hierarchical LLM architectures discussed in the concluding section.

KPIs	Relevancy	Correctness	Faithfulness
Vanilla Llama	70.3%	54.3%	42.2%
Enwar	81.2%	76.9%	68.6%

TABLE 1. KPI comparison for the scene in Fig. 2.

KPI Evaluation of State-of-the-Art LLMs on Modality Combinations

This section evaluates the performance of various state-of-the-art LLMs across different modality combinations, which is presented in Fig. 3 and Fig. 4 and discussed in the following subsections.

Modality Combination Comparison

For single modality evaluations, the general trend across KPI performance shows GPS < LiDAR< Camera (CAM). GPS alone provides limited contextual information, resulting in the lowest scores, while CAM proves to be the most effective single modality, offering richer visual context significantly enhancing answer relevancy, correctness, and faithfulness.

When dual modalities are combined, the trends observed in the single-modality evaluations continue. Specifically, when Camera or GPS is paired with LiDAR, CAM+LiDAR consistently outperforms GPS+LiDAR across all KPIs. This reflects the stronger impact of visual data on the models' ability to generate contextually rich and accurate responses. As expected, the integration of all three modalities yields the highest performance across every KPI. The fusion of spatial, depth, and visual information allows the models to deliver the most comprehensive and accurate responses, further emphasizing the value of multi-modal data integration for advanced situational awareness.

LLM Type and Size Comparison

Across all modality combinations, larger models yield higher absolute KPI values, reflecting the benefits of parameter scaling in LLMs. However, the rate of improvement diminishes as model size increases, with performance gains saturating beyond mid-scale models. The largest architectures offer only marginal advantages over slightly smaller ones. Notably, Mistral-7b and LLaMA-8b perform comparably across all metrics and modalities, indicating similar efficiency and effectiveness despite their size differences.

Figure 4 reveals a noticeable observation: Despite the larger models providing better overall absolute KPIs), the efficiency (i.e., performance per billion parameters) of adding more parameters decreases significantly, potentially indicating overfitting and interesting research directions covered in the next section. Another promising way of inference latency reduction might be training baby language models (LMs) to operate directly on the sensory data at the edge to form local RAG to eliminate the need for intermediary steps of modality transformation and information extraction.

In terms of deployment feasibility, the observed KPI saturation beyond mid-sized models highlights that smaller LLMs offer a more favorable cost-to-performance trade-off for practical use. Given their lower memory and energy requirements, models such as Mistral-7b and LLaMA-8b can achieve near-real-time inference within moderate hardware budgets, making them better suit-

As a RAG-empowered multi-modal LLM framework, ENWAR can address some of the key challenges in next-generation networks by enabling situational aware network management through multi-modal perception.

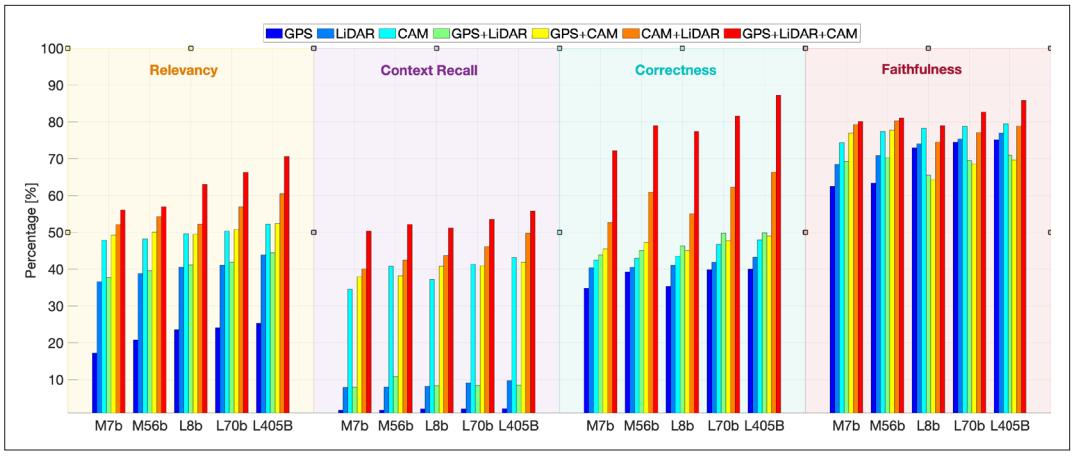

FIGURE 3. Absolute KPI [%] comparison of LLMs across modality combinations.

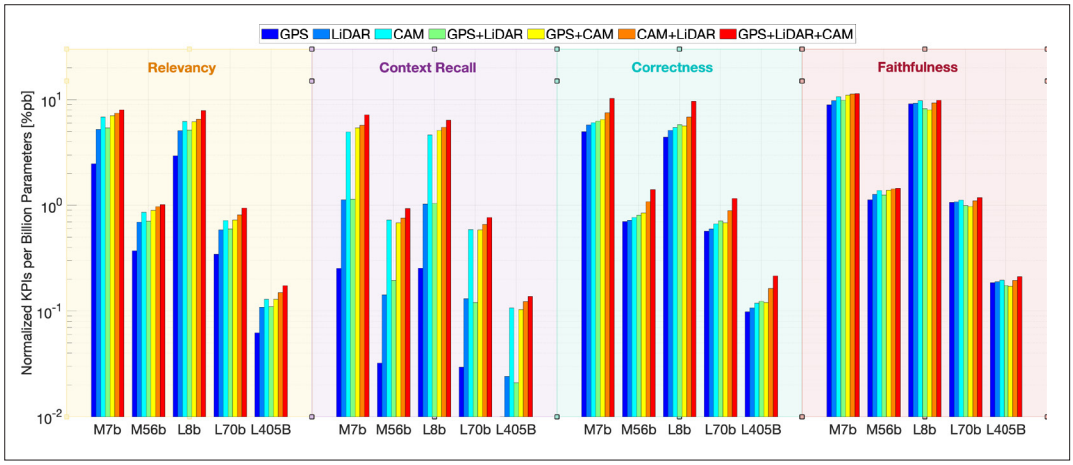

FIGURE 4. Normalized per billion parameter of each LLM's KPI [%pb].

ed for edge or distributed 6G deployments. This observation underscores that computational scalability in ENWAR dependent on model size and on efficient alignment between task complexity, available compute resources, and latency constraints.

Conclusion and Future Directions

As a RAG-empowered multi-modal LLM framework, ENWAR can address some of the key challenges in next-generation networks by enabling situational aware network management through multi-modal perception. By preprocessing and integrating various sensory data, ENWAR enhances its ability to interpret complex wireless environments and deliver contextually rich, human-interpretable insights. In spite of promising preliminary results, there is still room for improvement through several architectural enhancements depending on the target applications, which are discussed below.

Hierarchical and Federated LLM Architectures: For mission-critical and time-sensitive tasks, inference time and model efficiency can be significantly improved by adopting a federated LLM architecture that integrates smaller, edge-based "baby LM" with full-scale LLMs in the cloud. Baby LMs are designed for near-real-time operation at the edge, reducing reliance on cloud infrastructure. By employing model pruning and quantization techniques, these models remain lightweight and efficient, focusing on immediate, critical tasks. More complex computations are offloaded to cloud-based LLMs, providing both speed at the edge and scalability in the cloud. Further latency reduction can be achieved by training baby LMs to operate directly on the sensory data at the edge to form local RAG, bypassing intermediary steps of modality transformation and information extraction. We plan to implement these approaches on resource-constrained devices and AI accelerators to ensure their feasibility in practical deployments.

Cooperative and Adaptive RAG Formation: Given the central role of RAG in ENWAR, a distributed and adaptive approach can maintain a global knowledge base by aggregating localized ones across the hierarchical LLM structure. This cooperative system allows lightweight edge models to access relevant, up-to-date information without storing large datasets locally. Continuous adaptive learning refines ENWAR's perception of dynamic environments, ensuring efficient multi-modal processing and mitigating overfitting. By synchronizing global updates and optimizing resource allocation through serverless computing, this design balances performance with scalability, enabling real-time, context-aware reasoning under practical deployment constraints.

From Multi-Modal Perception to Intelligent Network Optimization: Building upon the perception and reasoning capabilities established in ENWAR, future extensions will focus on translating environment awareness into actionable control

within wireless networks. The multi-modal grounding and retrieval mechanisms introduced here can be coupled with traditional and learning-based wireless optimization techniques to address system-level tasks such as beam prediction, beamforming, and blockage mitigation. By leveraging agentic reasoning structures, the framework can dynamically orchestrate specialized models that integrate perception with network decision making, enabling context-driven responses to evolving channel and mobility conditions. This approach bridges the gap between semantic environment understanding and applied wireless network solutions, solidifying the way for intelligent, self-optimizing, and situation-aware 6G networks.

References

[1] A. Celik and A. M. Eltawil, "At the Dawn of Generative AI era: A Tutorial-Cum-Survey on New Frontiers in 6G Wireless intelligence," *IEEE Open J. Commun. Soc.*, vol. 5, 2024, pp. 2433–89.

[2] A. Vaswani *et al.*, "Attention is All You Need," *Adv. Neural Inf. Process. Sys.*, 2017.

[3] A. Alkhateeb, S. Jiang, and G. Charan, "Real-Time Digital Twins: Vision and Research Directions for 6G and Beyond," *IEEE Commun. Mag.*, 2023.

[4] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," ArXiv, vol. abs/2312.10997, 2023.

[5] S. Tarkoma *et al.*, "AI-Native Interconnect Framework for Integration of Large Language Model Technologies in 6G Systems," arXiv preprint arXiv:2311.05842, 2023.

[6] Z. Guo *et al.*, "VLM-Auto: VLM-Based Autonomous Driving Assistant with Humanlike Behavior and Understanding for Complex Road Scenes," *2024 2nd Int'l. Conf. Foundation and LLMs*, 2024, pp. 501–507.

[7] X. Zhou *et al.*, "Vision Language Models in Autonomous Driving: A Survey and Outlook," *IEEE Trans. Intelligent Vehicles*, 01 2024.

[8] J. Shao *et al.*, "WirelessLLM: Empowering Large Language Models Towards Wireless Intelligence," arXiv preprint arXiv:2405.17053, 2024.

[9] M. Xu *et al.*, "When Large Language Model Agents Meet 6G Networks: Perception, Grounding, and Alignment," *IEEE Wireless Comms.*, 2024.

[10] A. Abdallah *et al.*, "NetOrchLLM: Mastering Wireless Network Orchestration with Large Language Models," arXiv preprint arXiv:2412.10107, 2024.

[11] L. Bariah *et al.*, "Large Generative AI Models for Telecom: The Next Big Thing?," *IEEE Commun. Mag.*, pp. 1–7, 2024.

[12] S. Xu *et al.*, "Large multi-modal models (LMMs) as Universal Foundation Models for AI-Native Wireless Systems," *IEEE Network*, 2024.

[13] J. Morais *et al.*, "DeepSense-V2V: A Vehicle-to-Vehicle Multimodal Sensing, Localization, and Communications Dataset," ArXiv, vol. abs/2406.17908, 2024.

[14] W. Dai *et al.*, "InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning," *Int'l. Conf. Neural Information Processing Systems*, ser. NIPS '23, 2024.

[15] Z. Li *et al.*, "Towards General Text Embeddings with Multi-Stage Contrastive Learning," arXiv preprint arXiv:2308.03281, 2023.

Biographies

Ahmed M. Nazar (amnazar@iastate.edu) received a Ph.D. in computer engineering from Iowa State University (ISU), USA, and is a postdoctoral scholar at ISU.

Abdulkadir Celik received a Ph.D. in electrical and computer engineering from Iowa State University, USA, in 2016. He is an associate professor at the University of Southampton, U.K.

Mohamed Y. Selim received a Ph.D. in computer engineering from Iowa State University, USA, in 2018; and is an associate teaching professor at ISU.

Asmaa Abdallah received a Ph.D. in electrical engineering from the American University of Beirut, Lebanon, in 2020. She is a research scientist at KAUST.

Daji Qiao received a Ph.D. degree in Electrical Engineering from The University of Michigan, USA. He is a full professor at ISU.

Ahmed M. Eltawil received a Ph.D. degree in electrical engineering from the University of California, Los Angeles, USA, in 2003. He is a full professor at KAUST.