GLASS: Knowledge Graph-Guided for LLM-Assisted Semantic Communication

GLASS: Knowledge Graph-Guided for LLM-Assisted Semantic Communication

Loris Bacaloni, Alessio Sacco
Department of Control and Computer Engineering, Politecnico di Torino, Italy

***Abstract*—Conventional wireless communication systems are designed for exact bit-level delivery, but this objective has become inefficient in noisy, bandwidth-constrained environments and misaligned with modern applications that primarily require preservation of semantic meaning. While several semantic communication solutions have been proposed, for example, using deep learning–based joint source–channel coding, recent solutions mostly rely on Large Language Models (LLMs) for implicit semantic representations, but offering limited interpretability, controllability, and structured reasoning. In this paper, we present GLASS, an end-to-end semantic communication framework for text that shifts the objective from bit accuracy to meaning fidelity. Our key contribution is a hybrid transmission architecture that integrates Knowledge Graphs (KGs) with LLMs. The system extracts entities and relations to build a structured semantic representation, applies LLM-based semantic compression, and transmits compact meaning-aware representations over impaired wireless channels. We evaluate the framework under varying channel conditions using a standard text classification benchmark and demonstrate that GLASS significantly reduces transmitted data while preserving high semantic accuracy under channel degradation, showing that structured, meaning-centric communication is a robust and efficient alternative to traditional bit-level wireless systems.**

***Index Terms*—semantic communication, LLMs, knowledge graph**

I. Introduction

Traditional wireless communication systems, based on Shannon's information theory, are designed to ensure that every transmitted bit is received exactly as sent. This bit-level reliability is effective under stable channel conditions and sufficient bandwidth, but it becomes inefficient in practical environments characterized by noise, fading, interference, or limited spectral resources. In many emerging applications, such as Internet of Things (IoT) networks, unmanned aerial vehicles (UAVs), and edge computing systems, the critical requirement is not exact bit reconstruction, but the accurate delivery of the *meaning* of the message [1]–[4]. Semantic communication directly addresses this paradigm shift, moving the focus from bit accuracy to semantic fidelity [5]–[8]. Rather than transmitting every symbol, semantic systems encode and convey a compact representation of the underlying meaning, which is then reconstructed at the receiver using shared knowledge, contextual information, or learned language models. This approach can significantly reduce transmission overhead while preserving task-relevant information.

While several solutions have been proposed in this domain, ranging from deep learning–based joint source–channel coding to knowledge-aware encoding schemes, most approaches rely primarily on implicit semantic representations learned by neural networks or Large Language Models (LLMs), with limited interpretability and controllability.

In this paper, we introduce GLASS, a framework that integrates a Knowledge Graph (KG) with a Large Language Model (LLM) to enable efficient and interpretable semantic communication. The KG explicitly structures entities and relationships, grounding messages in a shared semantic space and allowing the system to transmit only the most relevant information, thereby improving efficiency, robustness, and semantic fidelity under varying channel conditions. Within this architecture, the KG captures the core semantic structure of the message, while the LLM handles natural language understanding, encoding, and reconstruction. Together, they enable accurate and bandwidth-efficient transmission of meaning, even in the presence of noise or resource constraints.

At the transmitter, a natural language processing pipeline extracts key entities and relations from the input text, producing a set of triples that form the KG representation. A sequence-to-sequence encoder (*e.g.,* T5 [9] or BART [10]) then performs semantic compression, converting the content into a compact token sequence with contextual embeddings. The encoded representation is received and processed by a two-stage semantic decoder. First, the same LLM employed at the transmitter produces a preliminary reconstruction from the potentially corrupted embeddings. Subsequently, a BERT-based masked language model refines the output by correcting uncertain tokens through contextual reasoning, thereby enhancing semantic consistency and robustness.

We evaluate our system on the SST-2 sentiment analysis dataset [11], under varying signal-to-noise ratios (SNRs) to simulate channel noise, assessing both semantic reconstruction quality and robustness to transmission errors. Results demonstrate that GLASS significantly reduces transmitted data size while preserving high semantic fidelity under channel degradation. We also experienced how T5 performs well as a semantic compressor, whereas BART often yields higher reconstruction quality when guided by KG-derived summaries. These findings support the premise that transmitting meaning—rather than raw bits—enables a more efficient, resilient, and context-aware communication paradigm.

The remainder of the paper is organized as follows. Section II reviews the state-of-the-art, and Section III overviews the proposed system. Section IV details the implementation and reports experimental results. Finally, Section V concludes the paper.

II. Related Work

Early work in semantic communication explored deep learning–based encoders and decoders that preserve task-relevant meaning rather than exact symbol reconstruction. Han et al. [12] introduced a semantic speech transmission system using a deep neural transceiver, demonstrating robustness to channel noise. Extending semantic communication to the visual domain, Lokumarambage et al. [13] proposed an end-to-end image transmission framework in which convolutional encoders extract high-level semantic features and a GAN-based decoder reconstructs realistic images at the receiver. These studies established semantic representation learning as a modality-agnostic paradigm across speech, text, and images.

LLM-Enabled Semantic Transmission. With the advent of transformer architectures, *Large Language Models (LLMs)* have emerged as powerful tools [14] for semantic communication, particularly for textual messages. Wang et al. [15] first demonstrated the feasibility of integrating LLMs directly into semantic communication systems by replacing traditional source and channel coding with prompt-based semantic encoding and decoding. Their work showed that LLMs can function as implicit semantic transceivers. Building on this idea, Chen et al. [16] proposed a hybrid semantic communication architecture combining LLMs with structured knowledge and retrieval-augmented generation (RAG). Messages are encoded as semantic triples and reconstructed at the receiver using LLM-based generation guided by external knowledge, improving contextual consistency and personalization. To address deployment constraints, Kalita [17] investigated LLM-based semantic communication in edge and IoT environments, proposing lightweight LLM deployment at edge nodes to reduce communication overhead. More recently, Salehi et al. [18] presented a fully LLM-enabled end-to-end semantic communication pipeline that models transmission as a continuous semantic transformation, focusing on intent preservation rather than literal reconstruction.

Knowledge Graph Integration. Parallel to LLM-based approaches, *Knowledge Graphs (KGs)* have been introduced to enhance semantic grounding and robustness. Jiang et al. [19] proposed an early KG-based semantic communication framework that transmits subject–predicate–object triplets with priority allocation based on semantic importance and employs KG reasoning for reconstruction at the receiver. Subsequent works explored deeper integration of structured knowledge. Guo et al. [20] proposed an LLM-driven KG construction framework for task-oriented semantic communication, enabling automated entity and relation extraction and dynamic graph updates. In contrast, Wang et al. [21] introduced a knowledge-enhanced receiver that retrieves factual triples from external knowledge bases to infer missing or corrupted semantic content. Liang et al. [22] further proposed KG-SemCom, which fuses contextual embeddings with KG entities to improve semantic recovery under low-SNR conditions.

Our Contribution. Motivated by these advances, this work proposes a unified semantic communication framework that jointly leverages KG-based semantic grounding and LLM-driven encoding and decoding. Unlike prior token-only or symbolic systems, the proposed architecture supports both token-based and embedding-based transmission modes, enabling graceful degradation across SNR regimes. Furthermore, a hybrid triplet extraction pipeline that combines Stanford OpenIE with a spaCy-based fallback improves robustness in knowledge extraction. This design yields a bandwidth-efficient and resilient semantic communication system suitable for noisy and resource-constrained wireless environments.

III. System Design

At a high level, GLASS transforms an input natural-language sentence into a compact semantic representation, transmits it over a noisy wireless channel, and reconstructs it at the receiver with minimal loss of meaning. The architecture is organized into three main processing phases: *Phase 1 - Semantic preprocessing and knowledge extraction*, which analyses the input sentence, extracts structured semantic triples, and builds a sentence-level knowledge graph; *Phase 2 - LLM-based semantic encoding and channel mapping*, which produces a compressed semantic representation using a transformer encoder-decoder model and maps it either to discrete tokens or continuous embeddings for transmission over the wireless channel; *Phase 3 - Contextual decoding and semantic refinement*, which reconstructs the sentence at the receiver using an LLM-based decoder and a BERT-based refinement stage for semantic consistency.

These three phases are implemented as modular software components that can be configured, enabled, or bypassed independently. A high-level block diagram of the proposed architecture is shown in Fig. 1. The diagram highlights the data flow between the semantic preprocessing block via a knowledge graph (KG), the LLM-based encoder, the wireless channel, and the semantic decoder.

The semantic preprocessing and KG construction block (*Phase 1*) operationalizes the notions of entities, relations, and graph-based semantic representations. KG provides a structured representation of knowledge as entities and relations, typically organized as labeled graphs or sets of triples [23]. When combined with relation extraction techniques, KGs offer a natural way to convert unstructured text into machine-interpretable semantic structures. In this context, they act as an intermediate layer that distills the core factual and relational content of a message before transmission.

To build KGs from raw text, it is necessary to extract relational tuples directly from natural-language sentences. In GLASS we use Open Information Extraction (OpenIE) [24] to extract relation tuples, typically binary relations, from plain text. Unlike traditional relation extraction, which targets predefined sets of relation types and requires annotated training data for each relation type, OpenIE aims to discover arbitrary relations in a domain-independent way. This technique is further combined with entropy-based sentence analysis to grasp the meaning of the sentence.

The LLM-based encoding block (*Phase 2*) instantiates the transformer encoder-decoder models. Transformers have be-

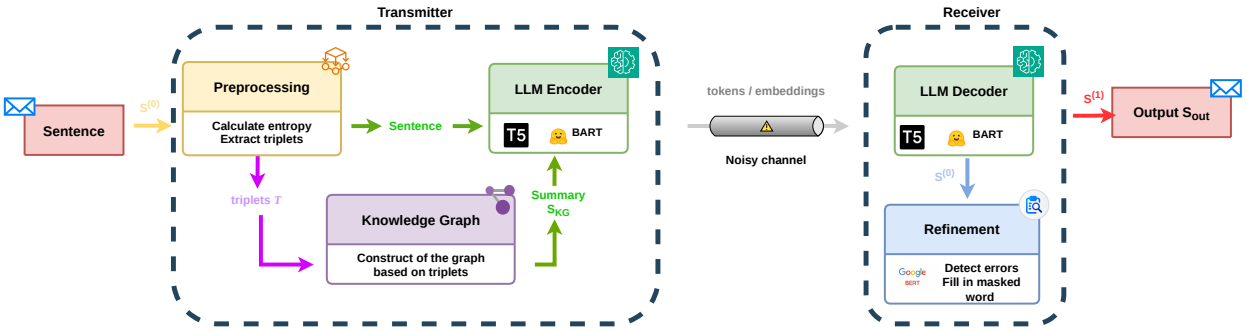

Fig. 1: High-level block diagram of GLASS. The transmitter combines KG and LLM for semantic communication, and the receiver decodes the message and refines possible missing words to reconstruct the original message.

come the dominant architecture for natural language processing, thanks to their ability to model long-range dependencies through self-attention and to support highly parallelizable training.In semantic communication, transformer-based sequence-to-sequence (seq2seq) models provide a natural mechanism for encoding text into compact semantic representations and reconstructing it at the receiver. Thanks to our Phase 1, additional cues derived from the KG are included in the input. The resulting compressed representation is then mapped to either a sequence of discrete tokens or a sequence of continuous embeddings.

In the context of semantic communication, physical-layer impairments result in distortions of the representation being transmitted: discrete symbols corresponding to tokens, or continuous-valued samples corresponding to embedding components. The semantic decoding and refinement block (*Phase 3*) can address these issues and implement contextual decoding strategies, combining LLM-based reconstruction with BERT-based masked language modeling into a single receiver-side pipeline. In our design, semantic decoding is performed in two steps, both based on transformer models: *(i)* an *LLM-based semantic decoder* (T5 or BART) that maps corrupted tokens or embeddings produced by the wireless channel into an initial textual reconstruction; *(ii)* a *refinement module* based on a masked language model (BERT) that detects and corrects local inconsistencies while preserving the global semantics [25].

The system design is guided by a set of assumptions that reflect the targeted application scenarios and the practical limitations of the underlying models. First, the framework focuses on short-to-medium-length sentences, as typically found in sentiment analysis or intent classification datasets, rather than on long documents. This choice aligns with the computational cost of transformer architectures and the need to maintain low latency in wireless communication settings. Second, the current implementation assumes English text and uses pre-trained models primarily trained on English corpora. Extending the framework to multilingual scenarios would require appropriate multilingual LLMs and potentially language-specific adaptations of the semantic preprocessing pipeline. Third, the wireless channel module is designed to emulate bandwidth-limited and noisy conditions, with configurable signal-to-noise ratios (SNRs) and transmission modes. Rather than modeling a specific physical deployment, the channel serves as a flexible abstraction that allows us to stress-test the semantic pipeline under varying levels of degradation. Finally, the framework is built around pre-trained transformer models and does not rely on task-specific fine-tuning of large models, thereby keeping computational requirements compatible with edge or resource-constrained deployments. Where necessary, smaller variants of the base models can be selected, and the compression strength can be adjusted through decoding parameters such as maximum length or beam size. The results underline that pre-trained models achieve satisfactory performance, but custom-tailored models can further improve reconstruction performance on specific application tasks.

A. *End-to-End* GLASS *Semantic Communication*

GLASS consists of three main phases of: semantic analysis and KG construction at the transmitter; LLM-based semantic compression and channel mapping; and receiver-side reconstruction with contextual refinement. More formally, in semantic communication, a sentence S is associated with a random variable M representing the meaning or task-relevant information associated with the sentence [26]. The mapping $\pi : S \mapsto M$ may correspond to a latent embedding, a set of logical propositions, or a downstream task label. At the receiver, the objective is to recover $\hat{M}$ rather than the exact sentence S.

A semantic distortion measure is defined as

$$d_{\text{sem}}(M, \hat{M}) = 1 - \sigma_{\text{sem}}(M, \hat{M}), \tag{1}$$

where $\sigma_{\text{sem}} \in [0, 1]$ denotes a semantic similarity metric, such as embedding similarity, symbolic overlap, or task accuracy. The design objective is

$$\min_{f,g} \mathbb{E}\left[d_{\text{sem}}(M, \hat{M})\right], \tag{2}$$

subject to bandwidth, power, or latency constraints [27]. Notably, semantic distortion differs from technical distortion: low bit error rate does not ensure semantic fidelity, and semantically equivalent sentences may differ significantly at the symbol level.

In the following, we elaborate on the three phases of the proposed methodology, describing their design principles, operational mechanisms, and interactions.

B. Phase 1 - Semantic Preprocessing and Knowledge Extraction

In this phase, the transmitter analyzes the input sentence S to derive both lexical and structured semantic information. An entropy-based metric characterizes the lexical variability of S and guides subsequent processing. A hybrid extraction pipeline combining Open Information Extraction (OpenIE) and a spaCy [28]-based dependency parser is used to extract subject–predicate–object triples, which are merged and filtered to construct a sentence-level KG $G(S)$. When applicable, this graph is further used to generate a compact KG-driven textual summary that captures the core semantic content of the sentence. A KG is a labeled graph in which nodes represent entities (*e.g.,* persons, locations, or abstract concepts) and edges encode semantic relations among them (*e.g.,* bornIn, locatedIn, hasSentiment). Formally, a KG is defined as the tuple

$$G(S) = (\mathcal{V}, \mathcal{R}, \mathcal{T}), \tag{3}$$

where $\mathcal{V}$ denotes the set of entities, $\mathcal{R}$ the set of relation types, and $\mathcal{T}$ the set of triples satisfying

$$\mathcal{T} \subseteq \mathcal{V} \times \mathcal{R} \times (\mathcal{V} \cup \mathcal{L}), \tag{4}$$

with $\mathcal{L}$ representing literal values (*e.g.,* numbers, strings, or dates). Each triple $(s, p, o) \in \mathcal{T}$ corresponds to a directed edge from subject s to object o labeled by relation p. In practice, OpenIE addresses this task by identifying domain-independent relations without relying on a predefined schema [24].

Given a corpus $\mathcal{C}$ of sentences $s \in \mathcal{C}$, an OpenIE system produces a set of relational triples

$$\mathcal{T} = \bigcup_{s \in \mathcal{C}} \mathrm{OIE}(s), \tag{5}$$

where $\mathrm{OIE}(s)$ returns one or more tuples (s_i, p_i, o_i) capturing relations expressed in s. Early approaches relied on shallow syntactic patterns, whereas more recent systems incorporate dependency parsing, semantic role labeling, and neural models [29]–[31]. A typical OpenIE pipeline includes: *(i)* preprocessing (*e.g.,* tokenization and syntactic analysis), *(ii)* argument identification, *(iii)* relation phrase detection, and *(iv)* tuple construction with confidence scoring. In this work, OpenIE serves as a scalable mechanism for extracting subject–predicate–object triples from arbitrary text, which are subsequently used for the KG construction and semantic processing. Industrial NLP libraries such as spaCy [28] provide efficient tools for analyzing text, including part-of-speech tagging, named entity recognition, and dependency parsing. Dependency parsing represents sentence structure as a tree, enabling the extraction of subject–predicate–object relations using simple grammatical patterns, such as verb-centered (subject/object), copular (is/are + attribute), and prepositional constructions.

In GLASS we combine OpenIE and spaCY: OpenIE offers broad coverage, while syntactic rules refine or correct extracted tuples. This hybrid strategy improves robustness, particularly for noisy or structurally complex text common in real-world semantic communication scenarios.

C. Phase 2 - LLM-Based Semantic Encoding

The semantic input S_{sem} (either the KG-based summary or the original sentence when KG extraction is bypassed) is passed to an LLM that performs semantic compression. The encoder maps S_{sem} to a shorter, context-rich representation $S_{\text{enc}} \in \mathbb{R}^{n \times d}$, which can be transmitted over the wireless channel in either token-based or embedding-based form, enabling flexible trade-offs between robustness and efficiency. In this work, we consider two widely used sequence-to-sequence (seq2seq) transformers, T5 [9] and BART [10]. T5 adopts a unified "text-to-text" framework, in which all tasks are formulated as mapping an input text sequence to an output text sequence via an encoder–decoder architecture. In contrast, BART combines a bidirectional transformer encoder with an autoregressive decoder and is pre-trained as a denoising autoencoder, where corrupted input sequences (*e.g.,* via masking, deletion, or permutation) are reconstructed. Transformers can be used to reproduce two alternative representations relevant to semantic communication:

1) **Token-based:** a discrete output sequence $\mathbf{Z} = (y_1, \ldots, y_m)$;
2) **Embedding-based:** continuous decoder states $\mathbf{H}_{\text{dec}} \in \mathbb{R}^{m \times d}$.

Transmitting $\mathbf{Z}$ corresponds to sending vocabulary indices that can be encoded using conventional source and channel coding. In contrast, transmitting $\mathbf{H}_{\text{dec}}$ requires quantization or analog mapping of real-valued vectors, leading to different robustness and bandwidth trade-offs. Token-based representations are discrete and interpretable but sensitive to bit errors, which can lead to uneven semantic degradation and require error control or refinement. In contrast, embedding-based representations are continuous and typically more robust to small perturbations, with distortion measured using metrics such as MSE or cosine similarity, albeit at the cost of higher dimensionality and reduced interpretability. The selection depends on bandwidth constraints, interpretability requirements, and robustness considerations. The comparative performance of these two strategies is evaluated and discussed in Section IV.

D. Phase 3 - Semantic Decoding and Contextual Refinement

Finally, the receiver reconstructs a fluent and semantically faithful sentence from the noisy channel output. Depending on the selected transmission mode, decoding operates on corrupted token sequences or contextual embeddings. An initial LLM-based reconstruction is followed by a refinement stage that mitigates noise-induced artifacts and optionally leverages the KG in Phase 1 to improve semantic consistency. The LLMs, such as T5 and BART, implement a conditional seq2seq mapping $p_{\theta_{\text{dec}}}(y \mid u)$. In the proposed framework, the decoder operates on a noisy channel output $\hat{Z}$ rather than on the clean source representation. In *token-based mode*, $\hat{Z}$ is a corrupted token sequence. In *embedding-based mode*, $\hat{H} \in \mathbb{R}^{m \times d}$ consists of continuous encoder states perturbed by the wireless channel. In both modes, the decoder produces an initial reconstruction $S^{(0)}$ that aims to preserve the original meaning despite channel distortions. However, such initial reconstruction $S^{(0)}$ may

contain residual errors. To enhance quality, we apply a second-stage refinement using BERT [25], pre-trained with a masked language modeling (MLM) objective. Given a sentence $\mathbf{w}$ and masked positions $\mathcal{M}$, MLM maximizes

$$\sum_{i \in \mathcal{M}} \log p_{\theta_{\text{mlm}}}(w_i \mid \mathbf{w}_{\setminus \mathcal{M}}), \tag{6}$$

enabling bidirectional context-based prediction. In our framework, likely erroneous tokens in $S^{(0)}$ are masked based on channel indicators or simple heuristics, and BERT predicts refined replacements, yielding the corrected sentence $S^{(1)}$.

Together, these phases transform a raw input sentence into a compact semantic representation that is resilient to wireless channel impairments and can be accurately reconstructed at the receiver.

IV. Evaluations and Results

In this section, we experimentally assess whether the GLASS pipeline can improve transmission efficiency and semantic robustness relative to purely symbolic baselines when messages are conveyed over a realistic radio channel with fading and additive noise. The experiments aim to quantitatively connect the "physical" layer (SNR and modulation scheme) with the semantic layer (quality of the reconstructed text) and with the compression gain provided by the proposed approach.

A. Experimental Setup

All experiments are conducted on the binary version of the *Stanford Sentiment Treebank* (SST-2) [11], using a fixed validation split of 872 short sentences annotated with positive or negative sentiment. The dataset is well aligned with our semantic communication setting: short inputs limit confounding effects from long-range dependencies, sentiment classification requires preserving fine-grained semantic cues (*e.g.,* polarity and negation), and SST-2 is a widely adopted benchmark that facilitates comparison with prior work. For each model configuration and channel condition, the full dataset is processed end to end, enabling direct comparisons of compression efficiency, channel robustness, and semantic fidelity across identical inputs.

We evaluate four semantic pipelines obtained by combining KG-based preprocessing (KG ON / KG OFF) with two encoder–decoder LLMs, namely T5 (`t5-small`) and BART (`bart-base`), yielding KG-T5, KG-BART, NoKG-T5, and NoKG-BART. In all cases, a BERT masked language model (`bert-base-uncased`) is employed at the receiver for semantic refinement. KG-enabled configurations perform semantic preprocessing and graph-based compression before LLM encoding, while baselines (vanilla LLMs) operate directly on the original sentence, enabling us to isolate the contribution of KG grounding. Decoding parameters (beam size, length penalty, repetition constraints) are kept fixed across all experiments to ensure fairness.

Two transmission modes are considered. In *token-based transmission*, discrete tokens produced by the decoder are transmitted as 24-bit units (16-bit token ID plus CRC), modulated using QPSK, and sent over a Rayleigh flat-fading channel with AWGN. In *embedding-based transmission*, continuous decoder representations are quantized to int8, transmitted as flat vectors over the same channel, and reconstructed prior to decoding. Both modes are evaluated over a common SNR grid $\{2, 4, 6, 8, 10\}$ dB using identical channel parameters.

Performance is assessed using channel-level metrics (BER, TER, or equivalent BER), text-level and semantic metrics (BLEU, ROUGE-L, METEOR, BERTScore, and SBERT similarity), and compression statistics (characters, tokens, transmitted kilobytes). Each configuration is evaluated across multiple runs to reduce variance caused by channel noise and decoding stochasticity. This experimental design enables a systematic comparison of KG-based preprocessing versus vanilla LLMs, token versus embedding transmission, T5 versus BART decoders, and robustness across SNR regimes, under consistent physical-layer conditions.

B. Wireless Channel Simulation

The wireless channel simulation block corresponds to the fading and noise models. In the context of semantic communication, physical-layer impairments translate into distortions of whatever representation is being transmitted: discrete symbols corresponding to tokens, or continuous-valued samples corresponding to embedding components. In this paper, we use the standard flat Rayleigh fading plus AWGN model, which is widely adopted in wireless communication theory [32], [33]. It injects Rayleigh fading, additive white Gaussian noise, and, when configured, multipath effects into the transmitted semantic representation. In a flat fading channel, the received baseband symbol at time index k is $y_k = hx_k + n_k$, where x_k and y_k denote the transmitted and received symbols, $h \in \mathbb{C}$ is the fading coefficient, and n_k is complex AWGN with variance N_0. This model is agnostic to the origin of x_k and thus applies to both discrete and continuous semantic representations.

Given our transformer-based semantic communication, for *token-level signals*, discrete tokens are mapped to modulation symbols (*e.g.,* PSK/QAM), and channel impairments cause symbol errors, while for *embedding-level signals*, continuous hidden states are transmitted as real or complex symbols, with channel noise modeled as additive perturbations in a high-dimensional vector space. While both cases experience channel-induced distortion, token transmission leads to combinatorial symbol errors, whereas embedding transmission results in geometric perturbations whose semantic impact depends on the structure of the embedding space and the robustness of the decoder.

C. Evaluation Metrics

We evaluate the proposed KG–LLM semantic communication framework using three complementary classes of metrics: *channel-level metrics* to validate the physical-layer behavior of the simulated link, *textual and semantic metrics* to assess meaning preservation after transmission and reconstruction, and *compression and bandwidth metrics* to quantify transmission efficiency.

TABLE I: Decoder performance vs. SNR, expressed as similarity metrics reported as percentages. The token-based version outperforms the alternatives.

SNR	Model	**Token-based**					**Embedding-based**				
		Sim.	ROUGE-L	METEOR	BLEU	BERT	Sim.	ROUGE-L	METEOR	BLEU	BERT
2 dB	KG-BART	**76.5**	**73.0**	**64.0**	**46.5**	**92.0**	73.0	64.0	61.5	44.0	89.8
	KG-T5	50.0	39.5	32.0	20.0	77.5	68.0	55.0	45.5	27.8	88.8
	NoKG-BART	60.5	61.0	57.5	34.5	88.5	**76.5**	**67.0**	**64.5**	**46.0**	**94.2**
	NoKG-T5	38.5	32.5	28.5	15.5	74.0	57.0	46.0	39.0	25.5	85.6
4 dB	KG-BART	**83.0**	**80.3**	**72.0**	**57.0**	**93.0**	76.5	69.0	66.0	49.8	90.8
	KG-T5	64.5	52.5	44.0	29.5	83.5	70.5	57.5	47.8	29.5	89.2
	NoKG-BART	66.8	67.0	66.0	42.5	89.8	**80.2**	**72.0**	**69.0**	**51.8**	**95.4**
	NoKG-T5	49.5	44.8	40.5	23.8	79.5	60.2	49.8	42.5	28.5	86.8
6 dB	KG-BART	**90.0**	**87.5**	**80.0**	**67.5**	**94.5**	80.5	73.8	70.0	55.5	92.0
	KG-T5	78.0	65.5	56.0	38.5	89.5	72.5	60.0	49.5	31.0	89.5
	NoKG-BART	72.0	73.0	73.0	50.5	91.5	**84.0**	**77.0**	**73.5**	**58.5**	**96.0**
	NoKG-T5	60.5	56.5	52.0	31.5	85.5	63.0	53.5	45.5	31.0	88.3
8 dB	KG-BART	**90.5**	**89.0**	**81.0**	**68.5**	**94.7**	81.2	75.0	71.0	56.8	92.2
	KG-T5	80.5	68.0	58.0	40.5	90.5	73.1	60.5	50.0	31.5	89.8
	NoKG-BART	72.5	73.5	73.0	51.0	91.3	**85.2**	**78.5**	**74.5**	**60.0**	**96.0**
	NoKG-T5	63.0	60.0	55.0	33.5	86.5	64.2	54.8	46.5	32.5	88.6
10 dB	KG-BART	**91.0**	**88.5**	**81.5**	**69.5**	**94.8**	81.8	76.0	72.0	58.5	92.4
	KG-T5	82.5	70.5	60.0	42.5	91.0	74.0	61.0	50.5	32.0	90.2
	NoKG-BART	72.5	73.5	73.0	51.5	91.5	**86.2**	**80.0**	**75.5**	**61.5**	**96.0**
	NoKG-T5	65.5	61.5	57.5	35.5	87.5	65.5	56.2	47.8	33.0	88.8

a) Channel-level metrics: For token-based transmission, we compute the *Bit Error Rate (BER)* as the ratio of incorrectly decoded bits to total transmitted bits, and the *Token Error Rate (TER)* as the fraction of corrupted tokens, where a token is marked erroneous if its ID is altered or the CRC check fails. BER and TER are used to verify that the simulated Rayleigh+AWGN channel exhibits the expected SNR-dependent behavior and to contextualize semantic performance under realistic physical-layer conditions. For embedding-based transmission, where no discrete bits are transmitted, we define an *equivalent BER* as a qualitative proxy derived from embedding distortion, using the mean squared error (MSE) and cosine similarity between transmitted and received embeddings. This indicator is not a physical bit-error probability, but is used solely to visualize how channel distortion evolves with SNR in embedding mode relative to token mode.

b) Textual and semantic metrics: To quantify reconstruction quality, we employ a mix of surface-level and embedding-based metrics. BLEU, ROUGE-L, and METEOR measure n-gram overlap, sequence structure, and synonym-aware lexical similarity, respectively, capturing how much textual content is preserved after compression and channel noise. To directly assess semantic fidelity, we additionally use BERTScore (F1), which compares contextual token embeddings, and Sentence-BERT (SBERT) cosine similarity, which measures global semantic proximity between original and reconstructed sentences. Together, these metrics provide a comprehensive view of both syntactic accuracy and meaning preservation.

c) Compression and bandwidth metrics: Transmission efficiency is evaluated using character-level and token-level compression ratios, defined as the ratio of transmitted to original text length or token count, aggregated over the full validation set. We also estimate the effective bandwidth required for transmission. In token-mode, bandwidth is computed assuming 24 bits per token (16-bit ID plus CRC), including framing overhead. In embedding-mode, bandwidth is estimated from the size of int8-quantized decoder embeddings, proportional to the product of sequence length and embedding dimension. These metrics allow us to directly relate semantic quality to compression level and radio resource usage across transmission modes and model configurations.

D. Results and Analysis

Table I reports the evolution of BLEU, ROUGE-L, METEOR, BERTScoreF1, and SBERT sentence similarity as a function of the SNR for both token-based and embedding-based transmission. For each SNR and metric, the best model value is bolded. Across all metrics and configurations, semantic quality increases monotonically with SNR. At $2,\mathrm{dB}$, performance is substantially degraded, while at 8–$10,\mathrm{dB}$ the residual channel errors become rare, and decoder behavior, rather than the wireless link, becomes the dominant source of variability.

In the *token-based* setting, KG-based preprocessing consistently improves performance. For both T5 and BART, KG-LLM variants outperform their NoKG counterparts across the entire SNR range and for all five metrics. The gains are most pronounced at intermediate SNRs (4–$6,\mathrm{dB}$), where channel noise is still significant: KG-BART achieves notably higher BLEU, ROUGE-L, and METEOR scores than NoKG-BART, and KG-T5 similarly outperforms NoKG-T5. Sentence similarity and BERTScore F1 follow the same trend, confirming that KG-based grounding mitigates the semantic impact of

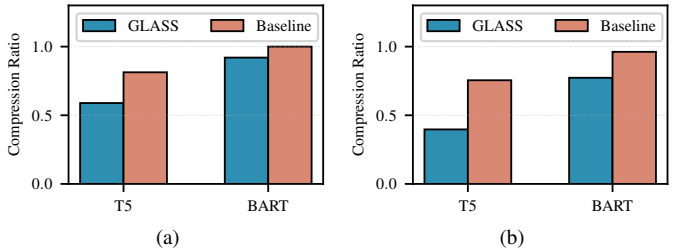

Fig. 2: Compression ratio of KG-LLM vs vanilla baselines (lower is better) for (a) character-level and (b) token-level scenarios.

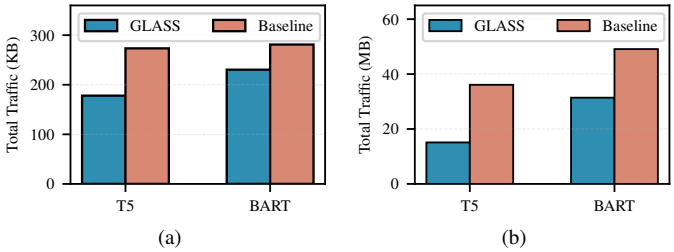

Fig. 3: Total traffic required by GLASS and vanilla models in (a) token-based and (b) embedding-based transmission.

token-level errors. Comparing T5 and BART in token-mode, KG-BART consistently achieves the highest semantic scores, reflecting BART's strong reconstruction ability stemming from its denoising pretraining. While KG-T5 is more aggressive in semantic compression (see Fig. 2), its scores remain slightly below those of KG-BART at comparable SNRs.

In the *embedding-based* setting, performance varies more smoothly with SNR. Because small perturbations in the embedding space do not necessarily cause catastrophic decoding errors, the curves exhibit gradual degradation and no sharp low-SNR cliff. For T5, KG-based preprocessing remains beneficial across all SNRs, particularly on semantic metrics such as sentence similarity and BERTScore, and also improves ROUGE-L and METEOR. BLEU is the only metric where NoKG-T5 matches or slightly exceeds KG-T5 at high SNR, reflecting BLEU's sensitivity to exact n-gram overlap rather than semantic equivalence. In BART's embedding mode, the effect of KG preprocessing is less pronounced. NoKG-BART often attains the highest scores, especially on BERTScore and sentence similarity, suggesting that when rich continuous representations are transmitted, BART can already exploit its internal semantic structure without explicit KG-based summarization. In this regime, overly aggressive compression may remove fine-grained details that would otherwise be preserved in the embedding stream. Nevertheless, KG-BART remains competitive and generally outperforms T5-based pipelines at the same SNR.

Overall, these results indicate that (i) KG-based preprocessing systematically improves robustness for T5 in both token- and embedding-based transmission; (ii) for BART, KG preprocessing provides substantial benefits in token-mode but a smaller, and occasionally negative, effect in embedding-mode; and (iii) embedding-based transmission degrades more gracefully with SNR, while token-based transmission can achieve slightly higher peak performance under sufficiently clean channel conditions.

We now measure the benefits in terms of reduced message size, aggregating statistics over the full SST-2 validation set of 872 sentences. Character-level and token-level compression ratios are reported in Fig. 2, while Fig. 3 summarizes the traffic saving in token-based and embedding-based modes.

As observed in Fig. 2a, GLASS achieves substantial compression compared to baselines (vanilla T5 or BART models). For T5, the compression ratio drops from approximately 0.81 (baseline) to 0.59 with GLASS, corresponding to a reduction of about 28% in the number of characters transmitted. For BART, the effect is milder but still noticeable, with the ratio going from 1.00 (no compression) to around 0.92, *i.e.,* roughly an 8% reduction. The effect is even more pronounced at the *token-level* in Fig. 2b. For T5, the token compression ratio decreases from about 0.76 in the baseline to 0.40 with KG-LLM, meaning that the number of tokens to be transmitted is almost halved (a reduction of 47.4%). For BART, the ratio drops from approximately 0.96 to 0.77, corresponding to a token reduction close to 20%. These results confirm that our Phase 1 significantly shortens the sequences seen by the encoder-decoder, especially for a summarization-oriented model such as T5.

These compression gains translate into tangible *bandwidth savings* at the physical layer (Fig. 3). In token mode (*i.e.,* QPSK, with 24 bits/token plus 280 bits of protocol overhead), the total traffic needed for the validation set decreases by $95.4\,\mathrm{kB}$ in T5 and $50.7\,\mathrm{kB}$ in BART. This corresponds to a bandwidth reduction of approximately one-third for T5 (34.9%) and nearly one-fifth (18.0%) for BART. In embedding mode (*i.e.,* int8 quantization with 1 byte per embedding value), we send more traffic, but the savings are larger in absolute terms. For T5, GLASS reduces the total transmitted volume by $16.2\,\mathrm{MB}$, *i.e.,* slightly more than 50%, while for BART, the reduction is of $12.9\,\mathrm{MB}$, yielding savings on the order of 26.4%. Since these quantities scale linearly with the number of messages, the impact on aggregate airtime and spectral efficiency becomes even more relevant in large-scale deployments. Taken together, the semantic performance curves and the compression–bandwidth tradeoff results indicate that KG-based preprocessing can significantly reduce the amount of information transmitted over the wireless link while preserving semantic fidelity. In conclusion, the results validate our approach. While the SST-2 dataset consists of short sentences, the pipeline is designed to operate on multi-clause inputs; however, the complexity of KG extraction and LLM latency may increase. We leave a systematic evaluation on longer corpora for future work.

V. Conclusion

This paper investigated whether a hybrid semantic communication architecture combining knowledge graphs (KGs) and large language models (LLMs) can transmit the meaning of

textual messages over noisy wireless channels more efficiently and robustly than conventional symbol-based approaches. By proposing GLASS, we leverage KG-based preprocessing to expose the core semantic structure of messages and reduce redundancy, while LLMs exploit this structured input to generate compact semantic representations and reconstruct fluent text at the receiver. End-to-end evaluations under noisy-channel conditions show that KG-aware configurations consistently preserve or improve semantic fidelity compared to LLM-only baselines, particularly at low-to-medium SNRs, while reducing transmission overhead. Token-based transmission is highly efficient under favorable channel conditions but degrades sharply as noise increases, whereas embedding-based transmission exhibits smoother performance degradation at the cost of different bandwidth trade-offs. Overall, the results demonstrate the viability of GLASS as a practical approach to bandwidth-efficient transmission in noisy environments.

REFERENCES

[1] S. Iyer, R. Khanai, D. Torse, R. J. Pandya, K. M. Rabie, K. Pai, W. U. Khan, and Z. Fadlullah, "A survey on semantic communications for intelligent wireless networks," *Wireless Personal Communications*, vol. 129, no. 1, pp. 569–611, 2023.

[2] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 2021.

[3] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2022.

[4] A. V. Ventrella, F. Esposito, A. Sacco, M. Flocco, G. Marchetto, and S. Gururajan, "Apron: An architecture for adaptive task planning of internet of things in challenged edge networks," in *2019 IEEE 8th International Conference on Cloud Networking (CloudNet)*. IEEE, 2019, pp. 1–6.

[5] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[6] D. Huang, X. Tao, F. Gao, and J. Lu, "Deep learning-based image semantic coding for semantic communications," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.

[7] Z. Zhang, Q. Yang, S. He, M. Sun, and J. Chen, "Wireless transmission of images with the assistance of multi-level semantic information," in *2022 International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, 2022, pp. 1–6.

[8] Y. Zhao, Y. Yue, S. Hou, B. Cheng, and Y. Huang, "LaMoSC: Large language model-driven semantic communication system for visual transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 10, no. 6, pp. 2005–2018, 2024.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[10] M. Lewis, Y. Liu *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 7871–7880.

[11] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2013, pp. 1631–1642.

[12] T. Han, Q. Yang, Z. Shi, S. He, and Z. Zhang, "Semantic-preserved communication system for highly efficient speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 245–259, 2022.

[13] M. U. Lokumarambage, V. S. S. Gowrisetty *et al.*, "Wireless end-to-end image transmission system using semantic communications," *IEEE Access*, vol. 11, pp. 37 149–37 163, 2023.

[14] A. Angi, A. Sacco, and G. Marchetto, "LLNet: An Intent-Driven Approach to Instructing Softwarized Network Devices Using a Small Language Model," *IEEE Transactions on Network and Service Management*, vol. 22, no. 4, pp. 3403–3418, 2025.

[15] Z. Wang, L. Zou, S. Wei *et al.*, "Large-language-model-enabled text semantic communication systems," *Applied Sciences*, vol. 15, no. 13, p. 7227, 2025.

[16] M. Chen, Z. Sun, X. He, L. Wang, and A. Al-Dulaimi, "Llm-based semantic communication: The way from task-originated to general," *IEEE Wireless Communications Letters*, vol. 14, no. 10, pp. 3029–3033, 2025.

[17] A. Kalita, "Large language models (llms) for semantic communication in edge-based iot networks," *arXiv preprint arXiv:2407.20970*, 2024.

[18] S. Salehi, M. Erol-Kantarci, and D. Niyato, "Llm-enabled data transmission in end-to-end semantic communication," in *2025 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2025, pp. 1–6.

[19] S. Jiang, Y. Liu, Y. Zhang *et al.*, "Reliable semantic communication system enabled by knowledge graph," *Entropy*, vol. 24, no. 6, p. 846, 2022.

[20] C. Guo, J. Liu, W. Gao *et al.*, "A large language model driven knowledge graph construction scheme for semantic communication," *Applied Sciences*, vol. 15, no. 8, p. 4575, 2025.

[21] B. Wang, R. Li, J. Zhu, Z. Zhao, and H. Zhang, "Knowledge enhanced semantic communication receiver," *IEEE Communications Letters*, vol. 27, no. 7, pp. 1794–1798, 2023.

[22] C. Liang, Y. Sun, D. Nyato, and M. Imran, "Knowledge graph fusion based semantic communication framework," *IEEE Transactions on Mobile Computing*, vol. 24, no. 11, pp. 11 416–11 429, 2025.

[23] A. Hogan, E. Blomqvist, M. Cochez, C. D'Amato *et al.*, "Knowledge graphs," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–37, 2021.

[24] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web," *Communications of the ACM*, vol. 51, no. 12, pp. 68–74, 2008.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[26] G. Xin, P. Fan, and K. B. Letaief, "Semantic communication: A survey of its theoretical development," *Entropy*, vol. 26, no. 2, p. 102, 2024.

[27] D. Wheeler and B. Natarajan, "Engineering semantic communication: A survey," *IEEE Access*, vol. 11, pp. 13 965–13 995, 2023.

[28] M. Honnibal and I. Montani, "spacy 101: Everything you need to know," https://spacy.io/usage/spacy-101, 2020, accessed: 2025-11-18.

[29] A. Yates, M. Banko, M. Broadhead, M. J. Cafarella, O. Etzioni, and S. Soderland, "TextRunner: Open information extraction on the web," in *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. Association for Computational Linguistics, 2007, pp. 25–26.

[30] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam, "Open information extraction: The second generation." in *IJCAI*, vol. 11, 2011, pp. 3–10.

[31] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 344–354.

[32] A. Goldsmith, *Wireless Communications*. Cambridge, UK: Cambridge University Press, 2005.

[33] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, UK: Cambridge University Press, 2005.