Attention-Driven Semantic Transmission Scheme for AI-Native Wireless Communications

Ki-Ho Lee, Hyun-Ho Choi, *Senior Member, IEEE*, and Jung-Ryun Lee, *Senior Member, IEEE*

***Abstract*— This letter proposes an attention-driven semantic transmission scheme for AI-native wireless communications. Leveraging self-attention scores from a pretrained bidirectional encoder representations from Transformers (BERT) model, the framework prioritizes semantically important tokens in both initial transmission and retransmission stages, without task-specific supervision. Experiments on sentences from public web sources confirm consistent improvements over conventional baselines in semantic fidelity, measured by cosine similarity and BERTScore. This model-agnostic approach provides a practical solution for robust and bandwidth-efficient communication, supporting future AI-native systems that prioritize meaning preservation over exact symbol reconstruction.**

***Index Terms*— Semantic communication, attention mechanism, BERT, hybrid ARQ.**

I. Introduction

As EMERGING services such as extended reality (XR), autonomous driving, and human-AI interaction demand communication systems that go beyond bit-level fidelity, semantic communication has recently emerged as a transformative paradigm. Unlike traditional methods that focus on exact reconstruction of transmitted symbols, semantic communication seeks to ensure that the receiver correctly interprets the intended message despite potential imperfections in channel conditions [1], [2]. A fundamental insight is that not all elements of a message contribute equally to its meaning. This motivates the notion of *content-importance awareness*, where critical components are identified and selectively treated to maintain semantic fidelity while maximizing efficiency.

Building on this perspective, recent studies have extended semantic communication to explicitly account for semantic importance and system heterogeneity. In particular, probabilistic semantic communication frameworks adapt compression ratios under rate-splitting multiple access (RSMA) to prioritize semantically critical information [3], while vision-based systems exploit task-specific gradient analysis or importance maps to identify essential features and transmit them with higher priority [4], [5]. Beyond importance-awareness, another line of work has focused on adapting semantic transmission under resource constraints: energy-efficient probabilistic semantic communication in space–air–ground integrated networks explores the trade-off between fidelity and energy consumption [6], and Swin Transformer-based dynamic frameworks address heterogeneous user capacities by allocating computing preferentially to essential semantic features [7]. More recently, unequal error protection (UEP) has been investigated in the context of semantic communications through multi-level reliability interfaces, where source and channel coders independently adapt protection levels to semantic importance [8].

Despite these advances, two key limitations remain unaddressed. First, there is no unified framework for quantifying and standardizing semantic importance across different tasks. In particular, although large-scale General AI models inherently provide pre-trained representations from which contextual importance signals can be simultaneously extracted, current communication-oriented approaches do not leverage this capability. The absence of such integration limits the ability to consistently reuse pretrained importance across diverse tasks and transmission settings. Second, conventional hybrid automatic repeat request (HARQ) protocols based on transport block (TB) transmission [9], [10] remain agnostic to semantic content. Existing HARQ mechanisms uniformly detect errors and trigger retransmissions without prioritizing semantic value, leading to inefficiencies, especially when certain words or tokens are more critical to comprehension. Third, earlier ARQ-based voice transmission schemes [11] faced latency challenges due to short packet durations and long round-trip times. With 5G reducing end-to-end delay to around 1 ms, such latency constraints are largely mitigated, enabling short voice packets to be transmitted as TBs without significant delay.

Motivated by these limitations, this letter proposes a semantic-aware transmission framework that jointly addresses both issues. First, instead of relying on handcrafted or probabilistic heuristics, we leverage a pretrained BERT model [12], [13] to estimate token-level semantic importance. Such pretrained language models capture contextual relevance beyond frequency or position, provide attention scores as interpretable indicators of importance, and generalize across domains without task-specific retraining. Since importance estimation relies on a pretrained model, the scores can be precomputed or cached in a higher layer, reducing the burden of processing during transmission. Second, based on these importance

Received 23 October 2025; revised 12 November 2025; accepted 23 November 2025. Date of publication 26 November 2025; date of current version 17 December 2025. This research was supported by the Chung-Ang University Research Grants in 2025 and in part supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea), through Robot Industry Core Technology Development Program (Development of shared autonomy control framework and AI-based application technology for enhancing tasks of hyper realistic telepresence robots in unstructured environment) under Grant 20023294. The associate editor coordinating the review of this letter and approving it for publication was Z. Yang. *(Corresponding author: Jung-Ryun Lee.)*

Ki-Ho Lee is with the School of Electrical Engineering, Chung-Ang University, Seoul 06974, South Korea (e-mail: kiholee79@cau.ac.kr).

Hyun-Ho Choi is with the School of ICT, Robotics and Mechanical Engineering, Hankyong National University, Anseong 17579, South Korea (e-mail: hhchoi@hknu.ac.kr).

Jung-Ryun Lee is with the School of Electrical Engineering and the Department of Intelligent Energy and Industry Convergence, Chung-Ang University, Seoul 06974, South Korea (e-mail: jrlee@cau.ac.kr).

Digital Object Identifier 10.1109/LCOMM.2025.3637247

scores, we design a semantic-aware HARQ framework that assigns UEP to tokens according to their semantic relevance. In this way, semantically vital tokens are mapped to more reliable channel resources, while less critical ones receive weaker protection. This importance-aware UEP scheme ensures that both initial transmission and retransmission resources are allocated where they matter most, thereby improving efficiency and fidelity beyond what existing frameworks can achieve.

II. Proposed Attention-Driven Semantic Transmission Scheme

A. Attention-Based Importance Estimation

We propose a token importance estimation scheme that leverages self-attention scores from a pretrained BERT model. BERT is chosen because its multi-head self-attention mechanism jointly encodes bidirectional context, allowing each token representation to reflect interactions with all other tokens in the sentence. This makes it particularly suitable for identifying contextually salient tokens. The underlying rationale is that tokens receiving consistently high attention from others tend to act as semantic anchors, often corresponding to syntactic heads or central content words. Recent studies [14] show that attention heads align with syntactic dependencies, while subsequent work [15] demonstrates that attention distributions highlight semantically meaningful content words. Taken together, these works support the use of attention scores as reliable proxies for token importance in the proposed transmission scheme.

To formalize this process, consider an input sentence tokenized as $X = \{x_1, x_2, \ldots, x_N\}$, where $x_i \in \mathcal{V}$ represents the i-th token in the sequence of length N. At the final layer L of the BERT encoder L, the hidden representation is expressed as $\mathbf{H}^{(L)} = [\mathbf{h}_1^{(L)}, \ldots, \mathbf{h}_N^{(L)}]^\top \in \mathbb{R}^{N \times d_{\text{model}}}$, where each $\mathbf{h}_i^{(L)} \in \mathbb{R}^{d_{\text{model}}}$ denotes the contextual embedding of token x_i. For each attention head $h \in \{1, \ldots, H\}$, the query, key, and value matrices are defined as

$$\mathbf{Q}_h^{(L)} = \mathbf{H}^{(L)} \mathbf{W}_{Q,h}^{(L)}, \ \mathbf{K}_h^{(L)} = \mathbf{H}^{(L)} \mathbf{W}_{K,h}^{(L)}, \ \mathbf{V}_h^{(L)} = \mathbf{H}^{(L)} \mathbf{W}_{V,h}^{(L)},$$

where $\mathbf{W}_{Q,h}^{(L)}, \mathbf{W}_{K,h}^{(L)}, \mathbf{W}_{V,h}^{(L)} \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are the projection matrices for queries, keys, and values, respectively. Here, d_{model} denotes the model (embedding) dimension (e.g., $d_{\text{model}} = 768$ in BERT-base), and $d_k = d_{\text{model}}/H$.

The scaled dot-product attention for head h is computed as

$$\mathbf{A}_h^{(L)} = \text{softmax}_{\text{row}}\left(\frac{\mathbf{Q}_h^{(L)} (\mathbf{K}_h^{(L)})^\top}{\sqrt{d_k}} + \mathbf{M} \right) \in \mathbb{R}^{N \times N}, \quad (1)$$

where $\mathbf{M}$ is an optional attention mask. The element $(\mathbf{A}_h^{(L)})_{i,j}$ denotes the attention weight from query token x_i to key token x_j. Each row forms a probability distribution:

$$\sum_{j=1}^{N} (\mathbf{A}_h^{(L)})_{i,j} = 1, \qquad (\mathbf{A}_h^{(L)})_{i,j} \geq 0. \quad (2)$$

Aggregating across heads, we obtain the mean attention matrix

$$\bar{\mathbf{A}} = \frac{1}{H} \sum_{h=1}^{H} \mathbf{A}_h^{(L)}, \qquad \bar{A}_{i,j} \triangleq (\bar{\mathbf{A}})_{i,j}. \quad (3)$$

TABLE I
Normalized Aggregated Attention Scores (Sample Sentence)

Token	Normalized aggregated attention score
the	0.1203
quick	0.0893
brown	0.0884
fox	0.0994
jumps	0.1220
over	0.0935
the	0.0987
lazy	0.0867
dog	0.0857

Finally, the *aggregated attention score* of token x_j is defined as

$$\tilde{A}_j = \frac{1}{N} \sum_{i=1}^{N} \bar{A}_{i,j}, \quad (4)$$

where a higher $\tilde{A}_j$ indicates stronger contextual centrality. This computation can be efficiently parallelized across multi-head structure, typically requiring only a few milliseconds per sentence on modern GPUs [16].

Table I presents sample normalized scores, where tokens such as `jumps` and the first occurrence of `the` exhibit high attention, reflecting their semantic or syntactic saliency. In contrast, tokens such as `lazy` and `dog` receive lower scores, suggesting a peripheral role. From a meaning-understanding perspective, the first `the` is crucial because it introduces the discourse subject, `the quick brown fox`, which represents the agent of the main action. Once this subject is established, subsequent references are interpreted relative to it, so attention naturally concentrates on this determiner as the entry point of the main proposition. In contrast, the second `the` occurs within a prepositional phrase (`over the lazy dog`), which serves an adjunct role rather than defining the sentence's primary event structure, and thus receives lower aggregated importance. Moreover, compared to other content words, the first `the` receives elevated attention because it anchors the main noun phrase that frames the sentence, making it indispensable for both syntactic parsing and semantic grounding. Likewise, the verb `jumps` attains high importance since it conveys the sentence's central action, whereas adjectives such as `lazy` naturally attract lower attention, reflecting their modifying rather than core semantic role.

B. Attention-Driven Initial Transmission

Given a target transmission rate $r = 1 - \text{TER}$, the number of transmitted tokens is $k = \lfloor r \cdot N \rfloor$, where $\lfloor \cdot \rfloor$ is the floor operation. The transmitted set is $\hat{X} = \{x_{i_1}, \ldots, x_{i_k}\}$, consisting of the k tokens with the highest scores $\tilde{A}_j$. This selection ensures that the most semantically significant tokens are prioritized.

To formalize importance-aware resource allocation, we define the protection level for token x_j as

$$P_j \propto f(\tilde{A}_j), \quad (5)$$

where P_j denotes the strength of channel protection (e.g., coding redundancy or power), and $f(\cdot)$ is a monotonic increasing function of the aggregated attention score $\tilde{A}_j$. In this way,

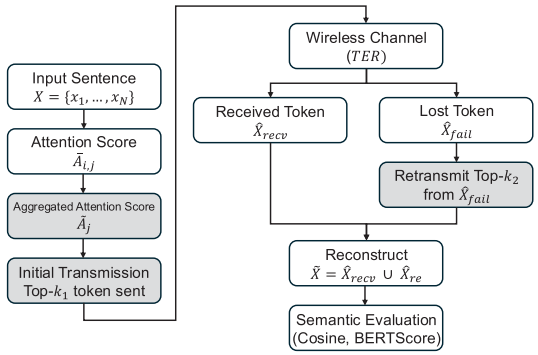

Fig. 1. Block diagram of the proposed semantic transmission scheme.

TABLE II
EXPERIMENTAL SETUP SUMMARY

Parameter	Value
Model	BERT-base (uncased, pretrained)
Attention Source	Final encoder layer
Vocabulary Size	30,522 ($b = 15$ bits/token)
Sentence Sources	Wikipedia, Wikinews
Number of Sentences	4,000 (2,000 from each source)
Sentence Length	5 – 15, 8 – 25 tokens
Channel Model	Rayleigh fading
SNR Levels	0 – 20 dB (step size 2 dB)
LDPC Block Length (n)	48
LDPC Code Rate (R)	$1/2$ ($k = 15 + 1$ (padding) + 8 (CRC))
Modulation	BPSK
Token Prioritization	Attention-based / Non-prioritized
Similarity Metrics	Cosine similarity / BERTScore

tokens with higher importance scores are always assigned stronger protection than those with lower scores. Unlike conventional equal treatment of tokens (non-prioritized), where each token is assigned the same constant protection level $P_j = C$ for $j = 1, \ldots, N$, this approach allocates protection levels proportionally to semantic relevance. Consequently, the transmitted subset $\hat{X}$ maintains the essential meaning of the original message while reducing bandwidth usage.

The proposed initial transmission procedure is described in the left half of Fig. 1. The first stage selects and protects the top-k_1 tokens based on attention scores, which are then sent through the wireless channel. To align with HARQ operations, each token is assumed to be transmitted as a single TB, meaning that every selected token corresponds to one coded packet occupying a complete TB unit.

C. Attention-Driven Retransmission

Let X denote the original sequence and $\hat{X}_{\text{recv}} \subset X$ the subset successfully received. The set of lost tokens is $\hat{X}_{\text{fail}} = X \setminus \hat{X}_{\text{recv}}$. The number of successfully received tokens is $k_1 = \lfloor (1 - \text{TER}) \cdot N \rfloor$, and the number available for retransmission is $k_2 = \lfloor (1 - \text{TER}) \cdot (N - k_1) \rfloor$.

The retransmission set is selected as

$$\hat{X}_{\text{re}} = \arg \max_{S \subset \hat{X}_{\text{fail}}, |S| = k_2} \sum_{x_j \in S} \tilde{A}_j, \tag{6}$$

where S contains the k_2 most important failed tokens.

As in the initial transmission, the retransmission stage also applies importance-aware protection, where each token $x_j \in \hat{X}_{\text{re}}$ is assigned a protection level $P_j \propto f(\tilde{A}_j)$. This contrasts with conventional HARQ, in which all tokens are retransmitted with equal protection ($P_j = C$ for all j). The reconstructed sequence is

$$\tilde{X} = \hat{X}_{\text{recv}} \cup \hat{X}_{\text{re}}, \tag{7}$$

with order restored according to the original X. The effective delivery rate becomes approximately $1 - \text{TER}^2$. This retransmission step corresponds to the right half of the block diagram in Fig. 1, where lost tokens are identified and the top-k_2 are selectively retransmitted to restore semantic fidelity.

III. EXPERIMENTAL SETUP

A. Experimental Procedure

We randomly sample a total of 4,000 sentences, equally drawn from Wikipedia and Wikinews. To examine the impact of sentence length on semantic transmission performance, we conduct experiments with two token length configurations: 5–15 tokens and 8–25 tokens (excluding special tokens such as [CLS] and [SEP]). For each sentence, token-level aggregated attention scores are computed from the mean attention weights in the final encoder layer of a pretrained BERT-base model. Each token is mapped to a payload of $b = 15$ bits, sufficient to represent all $30{,}522 (< 2^{15})$ vocabulary entries, which corresponds to the minimal identifier space needed for transmission-level symbolization. The 15-bit payload, together with an 8-bit cyclic redundancy check (CRC) and one zero-padded bit, forms a 24-bit block that is encoded using an LDPC encoder with $k = 24$ information bits per codeword. The encoded sequence is transmitted as a TB through a Rayleigh fading channel characterized by SNR, and the corresponding TER is obtained from LDPC-coded BPSK transmission and estimated via Monte Carlo simulations over multiple trials. The measured TER is subsequently applied to determine the number of retained tokens for both initial transmission and retransmission.

Although our experiment uses a single LDPC code with a fixed rate ($R = 1/2$, $n = 48$) for simplicity, a practical UEP system would utilize multiple channel codes with different protection levels to reflect token-wise priority. Instead of explicitly modeling multiple coding schemes, we approximate this behavior by assuming that higher-priority tokens are more likely to be successfully received; this abstraction captures the essential effect of importance-aware protection while avoiding system-level complexity, and its theoretical justification is provided in Appendix. Under this setup, we evaluate transmission quality by comparing the proposed attention-based token selection with a conventional non-prioritized semantic communication scheme, which randomly receives tokens within the given TER without employing UEP. The key experimental parameters are summarized in Table II.

B. Similarity Metrics

We evaluate semantic fidelity using two widely adopted embedding-based metrics: cosine similarity and BERTScore.

1) Cosine Similarity: Cosine similarity measures the angular closeness between two vector representations [17]. Here, it is applied to sentence-level embeddings computed by averaging token-level BERT embeddings.

Let $\mathbf{z}$ and $\tilde{\mathbf{z}}$ denote the original and reconstructed embeddings, respectively. The similarity is defined as

$$\text{Sim}_{\cos}(\mathbf{z}, \tilde{\mathbf{z}}) = \frac{\mathbf{z}^\top \tilde{\mathbf{z}}}{\|\mathbf{z}\| \cdot \|\tilde{\mathbf{z}}\|}. \tag{8}$$

A higher value indicates better global semantic preservation.

2) BERTScore: BERTScore provides a fine-grained token-level evaluation [18]. It uses contextualized BERT embeddings to measure the alignment between the original and reconstructed sentences. Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\tilde{X} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ be the original token embeddings and the reconstructed ones, respectively. The precision of BERTScore is defined as $P = \frac{1}{M}\sum_{j=1}^{M} \max_i \cos(\mathbf{y}_j, \mathbf{x}_i)$, and the recall as $R = \frac{1}{N}\sum_{i=1}^{N} \max_j \cos(\mathbf{x}_i, \mathbf{y}_j)$. The overall score is the harmonic mean of P and R, which is given as

$$\text{BERTScore}_{F_1} = \frac{2PR}{P+R}. \tag{9}$$

BERTScore effectively captures semantic overlap at the token level and is robust to paraphrasing and word-order variations.

IV. Results and Discussions

A. Semantic Similarity Evaluation for Initial Transmission

Fig. 2 compares the semantic similarity between the original and reconstructed sentences for two different number of tokens configurations. The proposed attention-based method shows better performance than the conventional non-prioritized scheme. Specifically, at 4 dB and 8–25 tokens, the attention-based cosine similarity and BERTScore reach approximately 0.82 and 0.92, respectively, while the non-prioritized method yields about 0.78 and 0.90. When comparing different number of tokens, the BERTScore shows negligible variation, indicating that semantic-level meaning is well preserved even with shorter token sequences. In contrast, cosine similarity exhibits a slightly larger gap at low-to-mid SNR, as shorter token sequences tend to yield less stable embedding alignment and higher feature-direction variance. This occurs because cosine similarity is computed from an averaged sentence-level embedding, where shorter sequences amplify token-level fluctuations, leading to larger directional variation in the latent space. This behavior implies that while semantic fidelity primarily depends on the correct transmission of key tokens, longer sequences help stabilize the embedding representation in the vector space. Additionally, the non-semantic successful initial transmission rate (1 - TER) as a reference shows that the semantic schemes surpass it.

Fig. 3 presents the semantic similarity results for retransmission. The successful transmission rate $1 - \text{TER}^2$ denotes the final rate after one retransmission in non-semantic communication. Compared with Fig. 2, the convergence toward unity occurs much earlier, achieving nearly perfect semantic reconstruction even at lower SNR values around 8 dB. This shows that retransmission plays an essential role in semantic communication, similar to the function of HARQ in conventional systems. The BERTScore again exhibits negligible variation regardless of token sequence length. However, as observed in the initial transmission results, cosine similarity remains more sensitive to sequence length since it is derived from the averaged embedding representation, where shorter sequences amplify token-level fluctuations. As a result, cosine similarity of shorter token sequences tends to produce less stable embedding alignment due to reduced contextual information. Nevertheless, retransmission effectively compensates for these variations by reinforcing high-importance tokens, which stabilizes the latent representation and preserves overall semantic robustness.

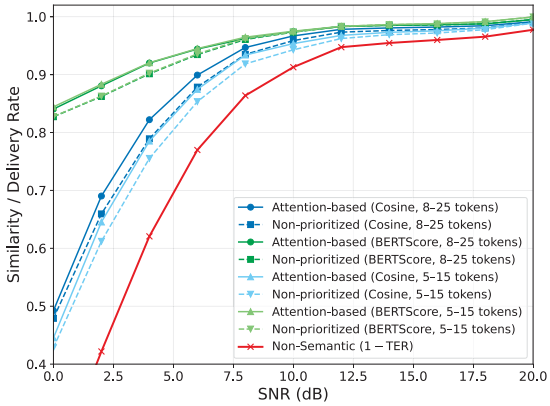

Fig. 2. Semantic similarity versus SNR for initial transmission.

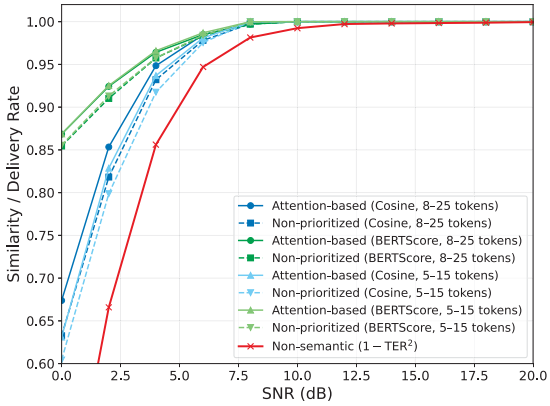

Fig. 3. Semantic similarity versus SNR for retransmission.

V. Conclusion

This letter proposed a semantic transmission framework that integrates token-level attention scores into both initial and retransmission stages to prioritize semantically important tokens. Performance gains were achieved over conventional baselines, confirming robust semantic preservation. The framework is model-agnostic and does not require task-specific supervision. For practical deployment, however, it is essential to maintain synchronization of the AI model between the transmitter and receiver to ensure consistent attention behavior and semantic interpretation, as even slight model deviations can lead to degradation in performance, which will be further investigated in future work. We will also implement UEP

using differentiated channel codes across attention-based token groups to achieve more accurate validation under identical transmission resources.

APPENDIX
MATHEMATICAL ANALYSIS OF THE UEP APPROXIMATION

To justify the idealized token-level reception assumption, we analyze a simplified two-class UEP model. Let N tokens be divided into two disjoint sets: the important set $\mathcal{I}$ with cardinality $|\mathcal{I}| = k_1$, and the non-important set $\mathcal{N}$ with $|\mathcal{N}| = N - k_1$. Each token x_j is decoded independently with success probability $s_H \in (0,1)$ if $x_j \in \mathcal{I}$ and $s_L \in (0,1)$ if $x_j \in \mathcal{N}$, where $s_H > s_L$. This abstraction captures a simple form of unequal reliability between high- and low-importance tokens.

We first examine the likelihood that exactly a subset $\mathcal{R} \subseteq \mathcal{X}$ of size k_1 is successfully received. Let $r \triangleq |\mathcal{R} \cap \mathcal{I}|$ denote the number of important tokens included in $\mathcal{R}$. Among all possible subsets $\mathcal{R}$, the probability $\Pr[\text{exactly } \mathcal{R} \text{ succeed}]$ is maximized when $\mathcal{R} = \mathcal{I}$, i.e., when all top-$k_1$ important tokens are received. For any other subset $\mathcal{R}$ missing $(k_1 - r)$ important tokens, the relative likelihood becomes

$$\frac{\Pr[\text{exactly } \mathcal{R} \text{ succeed}]}{\Pr[\text{exactly } \mathcal{I} \text{ succeed}]} = \left(\frac{(1-s_H)\, s_L}{s_H\,(1-s_L)}\right)^{k_1 - r} = \theta^{\,k_1 - r}, \tag{A.1}$$

where $\theta = \frac{(1-s_H)s_L}{s_H(1-s_L)}$, and $\theta \in (0,1)$. This indicates that every non-ideal configuration is suppressed by a multiplicative factor θ^{k_1-r}. The suppression increases exponentially with the number of misplaced tokens $(k_1 - r)$ and becomes more pronounced as the reliability gap $(s_H - s_L)$ widens.

To quantify this dependency, we express $\log(1/\theta)$ in terms of the *logit* function $\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right)$ as

$$\log \tfrac{1}{\theta} = \operatorname{logit}(s_H) - \operatorname{logit}(s_L) = \frac{s_H - s_L}{\xi(1-\xi)},$$

for some $\xi \in (s_L, s_H)$ by the mean-value theorem. Since $\xi(1-\xi) \le \frac{1}{4}$ for any $\xi \in (0,1)$, we obtain

$$\theta \le \exp\big(-4(s_H - s_L)\big), \tag{A.2}$$

which indicates a universal exponential suppression proportional to the reliability difference, independent of N or k_1.

To examine aggregate effects, consider the family of subsets that miss exactly m important tokens:

$$\mathcal{F}_m = \{\mathcal{R} : |\mathcal{R}| = k_1,\ |\mathcal{R} \cap \mathcal{I}| = k_1 - m\},$$

whose cardinality is $|\mathcal{F}_m| = \binom{k_1}{m}\binom{N-k_1}{m}$. Summing over all $\mathcal{R} \in \mathcal{F}_m$ yields

$$\sum_{\mathcal{R}\in\mathcal{F}_m} \frac{\Pr[\text{exactly } \mathcal{R} \text{ succeed}]}{\Pr[\text{exactly } \mathcal{I} \text{ succeed}]} \le \binom{k_1}{m}\binom{N-k_1}{m}\theta^m. \tag{A.3}$$

Normalizing over all configurations gives the compact lower bound

$$\Pr[\text{top-}k_1 \text{ tokens received}] \ge \frac{1}{1 + \sum_{m=1}^{k_1}\binom{k_1}{m}\binom{N-k_1}{m}\theta^m}. \tag{A.4}$$

Equations (A.1)–(A.4) together show that, under this two-class UEP abstraction, the configuration in which the top-k_1 important tokens are successfully received approximately dominates all competing outcomes. Inferior configurations are exponentially suppressed by θ, whose decay rate increases with $(s_H - s_L)$. Therefore, the idealized token-level reception assumption adopted in this work provides a close approximation to the reception behavior observed in practical multi-level UEP systems.

REFERENCES

[1] Q. Lan et al., "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–352, Dec. 2021.

[2] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 210–219, Feb. 2022.

[3] Z. Zhao et al., "Compression ratio allocation for probabilistic semantic communication with RSMA," *IEEE Trans. Commun.*, vol. 73, no. 9, pp. 7304–7318, Sep. 2025.

[4] J. Hu, F. Wang, W. Xu, H. Gao, and P. Zhang, "Scalable multi-task semantic communication system with feature importance ranking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[5] Y. Sheng, H. Ye, L. Liang, and S. Jin, "Semantic communication for cooperative perception based on the importance map," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Nov. 2023, pp. 177–182.

[6] Z. Zhao et al., "Energy-efficient probabilistic semantic communication over space-air-ground integrated networks," *IEEE Trans. Wireless Commun.*, vol. 24, no. 10, pp. 8814–8829, Oct. 2025.

[7] L. X. Nguyen et al., "Swin transformer-based dynamic semantic communication for multi-user with different computing capacity," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 8957–8971, Jun. 2024.

[8] T.-Y. Tung, H. Esfahanizadeh, J. Du, and H. Viswanathan, "Multi-level reliability interface for semantic communications over wireless networks," *IEEE Trans. Commun.*, vol. 73, no. 8, pp. 6023–6035, Aug. 2025.

[9] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Commun.*, vol. 49, no. 9, pp. 1473–1487, Sep. 2001.

[10] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 24, no. 5, pp. 630–643, May 2003.

[11] L.-J. Chen, R. Kapoor, K. Lee, M. Y. Sanadidi, and M. Gerla, "Audio streaming over Bluetooth: An adaptive ARQ timeout approach," in *Proc. 24th Int. Conf. Distrib. Comput. Syst. Workshops*, Feb. 2004, pp. 196–201.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[13] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2025, pp. 5998–6008.

[14] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? An analysis of BERT's attention," in *Proc. ACL Workshop BlackboxNLP, Analyzing Interpreting Neural Netw. (NLP)*, 2019, pp. 276–286 W19-4828.

[15] D. Jang, S. Byun, and H. Shin, "A study on how attention scores in the BERT model are aware of lexical categories in syntactic and semantic tasks on the GLUE benchmark," in *Proc. LREC-COLING*, 2024, pp. 1684–1689.

[16] Hugging Face. (2024). *Performance and Benchmarking: BERT-Base Inference Latency*. [Online]. Available: https://huggingface.co/docs/transformers/v4.17.0/en/benchmarks

[17] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.

[18] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–43.