Quest Journals
Journal of Software Engineering and Simulation
Volume 11 ~ Issue 3 (2025) pp: 07-16
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

An Entropy and QA-driven Method for Enhancing Factual Consistency in Abstractive Summarization

CongHao Cao [1], GuiChang Li[1], JianKang Han[1]
School of Control and Computer Engineering, North China Electric Power University, Baoding, 071051, China Corresponding Author: CongHao Cao

ABSTRACT: *Despite the significant progress in abstractive summarization techniques driven by pre-trained language models, factual inconsistencies between generated summaries and source documents remain a major challenge. These inconsistencies, often referred to as "hallucinations," can be categorized into two types: intrinsic (entity co-occurrence errors) and extrinsic (fabricated entities). To enhance the factual consistency of summaries, this paper proposes a novel two-stage correction framework—EQFCS (Entropy and Question-Answering-driven Factual Consistency Framework). First, a heuristic entropy-based method is used to identify hallucinated entities. Then, a question generation (QG) and question answering (QA) pipeline is employed for correction. Experiments on the XSum dataset demonstrate that this method significantly improves factual consistency metrics while preserving summary fluency and structure, outperforming existing mainstream approaches. This study provides an effective solution for factual correction in generative text summarization and offers new insights for improving summary quality and reducing hallucination phenomena in the future.*

KEYWORDS: *Abstractive Text Summarization, Factual Consistency, Hallucination Detection, Information Entropy*

R*eceived 22 Mar., 2024;* ***R****evised 28 Mar., 2025;* ***A****ccepted 04 Apr., 2025 © The author(s) 2025.*
P*ublished with open access at* ***www.questjournas.org***

I. INTRODUCTION

Factual consistency is a critical issue in abstractive text summarization, particularly in high-reliability scenarios such as news reporting, legal documentation, and medical information synthesis. With the rapid advancement of pre-trained language models (PLMs), generative summarization systems have achieved significant progress in fluency and coherence. However, factual errors—commonly referred to as "hallucinations"—remain a major challenge, limiting their trustworthiness and applicability in real-world tasks.

Hallucinations in summarization can be broadly categorized into two types: internal and external. Internal hallucinations occur when the generated summary contains entities or facts that exist in the source text but are incorrectly associated or combined. External hallucinations, on the other hand, involve information that is entirely absent from the source text. These hallucinations are often difficult to detect without referring to the original text and can significantly undermine the reliability of the generated summaries.

One of the fundamental causes of hallucinations lies in the uncertainty of the generation process. During decoding, generative models sample words based on probability distributions, which may reflect ambiguity or overconfidence, leading to incorrect outputs. Existing summarization models lack internal mechanisms to assess factual correctness during generation, making it difficult to prevent or correct hallucinations in real time.

To address this challenge, we propose a novel framework—Entropy and QA-driven Framework for Consistency in Summarization (EQFCS)—designed to enhance the factual consistency of abstractive summaries through a two-stage process. First, we analyze entropy values during the decoding process of a generative summarizer to detect potential hallucinated entities. High-entropy words typically indicate uncertainty in generation and are more likely to be hallucinated. Based on this observation, we use a dynamic entropy threshold to identify candidate hallucinated entities.

Second, we design a correction pipeline inspired by human cognitive behavior: when people are uncertain about a fact, they tend to ask questions and seek answers from reliable sources. Our method simulates

this process by generating targeted questions around suspected hallucinated entities and leveraging a question answering (QA) system to extract the correct information from the source document. The extracted information is then seamlessly integrated into the summary, preserving sentence structure while improving factual alignment.

Our experimental results on the benchmark XSum dataset demonstrate that the EQFCS framework significantly improves factual consistency metrics (FactCC and FactScore) without compromising summary fluency, as measured by ROUGE. Additionally, we provide a detailed analysis of entropy trends and hallucination patterns across different datasets to validate the effectiveness of our entropy-based detection mechanism.The main contributions of this paper are as follows:

We propose an entropy-based entity hallucination detection method that enables fine-grained identification of ptentially incorrect tokens during the summarization process.

We introduce a two-stage QA-based correction mechanism that simulates human reasoning, improving summary factuality while maintaining sentence fluency.

The remainder of this paper is organized as follows: Section 2 reviews related work on hallucination detection and summarization consistency. Section 3 provides a detailed introduction to the proposed EQFCS framework. Section 4 describes the experimental setup and results. Section 5 discusses the findings, concludes the paper, and suggests directions for future research.

II. RELATED WORK

2.1 TEXT SUMMARIZATION AND FACTUAL CONSISTENCY

2.1.1 RESEARCH PROGRESS IN ABSTRACTIVE TEXT SUMMARIZATION

The development of abstractive text summarization has been closely linked to advancements in neural network architectures. The field experienced rapid growth following the introduction of the Sequence-to-Sequence (Seq2Seq) framework by Sutskever et al[1]. in 2014, which provided a foundational structure for mapping input sequences to output sequences. Subsequently, Rush et al[2] (2015) proposed one of the earliest summarization models based on this framework, incorporating an attention mechanism to improve alignment between source texts and generated summaries.

Building on this foundation, subsequent models introduced various optimizations. For instance, bidirectional recurrent neural networks (BiRNNs)[3] were employed to better capture contextual dependencies within documents. Attention mechanisms were further refined to allow models to selectively focus on semantically important segments of the input text.

With the rise of Transformer-based[4] architectures, particularly after the success of BERT[5], researchers began leveraging large-scale pre-trained language models for summarization tasks. Models such as BART[6], T5[7], and PEGASUS[8] demonstrated strong performance by pretraining on denoising or reconstruction tasks and fine-tuning on summarization datasets. These models proved particularly effective in handling long input sequences and generating coherent summaries, marking significant progress in the field.

More recent research efforts have shifted toward enhancing factual consistency, reducing redundancy, and addressing exposure bias in training. Techniques such as hierarchical attention[9], contrastive learning[10], and reinforcement learning[11] have been explored to better align generated summaries with source content and human expectations.

2.1.2 CURRENT RESEARCH ON FACTUAL CONSISTENCY IN ABSTRACTIVE TEXT SUMMARIZATION

Owing to the inherent nature of generative models, their output process is difficult to control with precision, making the generated content susceptible to uncertainty and the inclusion of potentially fabricated information. In recent years, the issue of factual consistency in abstractive summarization has attracted considerable research attention.

Existing efforts aimed at improving factual consistency in generated summaries can be broadly categorized into the following three approaches:

(1) POST-EDITING APPROACHES FOR FACTUAL ERROR CORRECTION IN SUMMARIES

This line of research focuses on correcting hallucinated or inaccurate content after the summary has been generated. For example, Cao[12], Kryscinski[13], and others have proposed using large pre-trained language models to perform end-to-end factual correction of summaries. Dong et al[14]. introduced two strategies, SpanFact and Auto-regressive correction: the former progressively replaces hallucinated entities through entity masking and infilling, while the latter generates all replacement entities in an auto-regressive manner. Fabbri[15] et al. proposed a compression-based editing model that learns to remove incorrect entities while preserving factual information, thus improving both grammaticality and factual completeness. In further work, Dong et al[16]. also leveraged external knowledge graphs (e.g., Wikidata) to supplement and correct entity-level information in generated summaries, thereby enhancing entity-level consistency.

(2) FACT-AWARE ABSTRACTIVE SUMMARIZATION

This category of methods emphasizes incorporating structured factual information during the generation process. For instance, Cao et al[17]. proposed the FTSum model, which guides summarization with factual triples extracted via Open Information Extraction (OpenIE), encouraging the model to attend more closely to factual input. Gunel et al[18]. introduced entity representations from knowledge graphs as auxiliary information to improve the model's ability to distinguish factual content. Ryu et al[11]. proposed a multi-objective reinforcement learning approach that balances fluency with factual consistency. Song et al[19]. studied how training data affects hallucination tendencies, showing that careful selection of training samples can significantly improve factual accuracy. Choubey et al[20]. proposed an "anti-expert model" framework, which reduces hallucination likelihood through contrastive fine-tuning without increasing inference time. Dixit et al[21]. introduced a candidate ranking approach that selects summaries based on a trade-off between factual correctness and output diversity. Liu et al[22]. incorporated Natural Language Feedback (NLF) from humans to guide the model's generation behavior, improving its factual sensitivity and generation quality.

(3) TEXTUAL ENTAILMENT-BASED FACT VERIFIVATION

This class of approaches treats factual consistency as a textual entailment or natural language inference (NLI) problem. Zha et al[23]. proposed ALIGNSCORE, which evaluates the semantic alignment between the summary and source document to measure factual consistency. Falke et al[24]. combined beam search with entailment classifiers to select the summary candidate that best preserves logical alignment with the source. Roit et al[25]. further integrated reinforcement learning with NLI models to fine-tune the generation process, improving factual consistency scores through reward optimization.

III. METHOD

To address potential hallucinated vocabulary that may arise during the decoding process of generative text summarization models and to promptly correct factual errors in the summaries, this paper proposes a two-stage method based on information entropy and question-answering models. The method marks hallucinated entities by using the entropy values output by the generative model, with entities exceeding a certain threshold being flagged as hallucinations. It then uses a question generation model and a question-answering model to replace these hallucinated entities. The overall framework is shown in Figure (1).

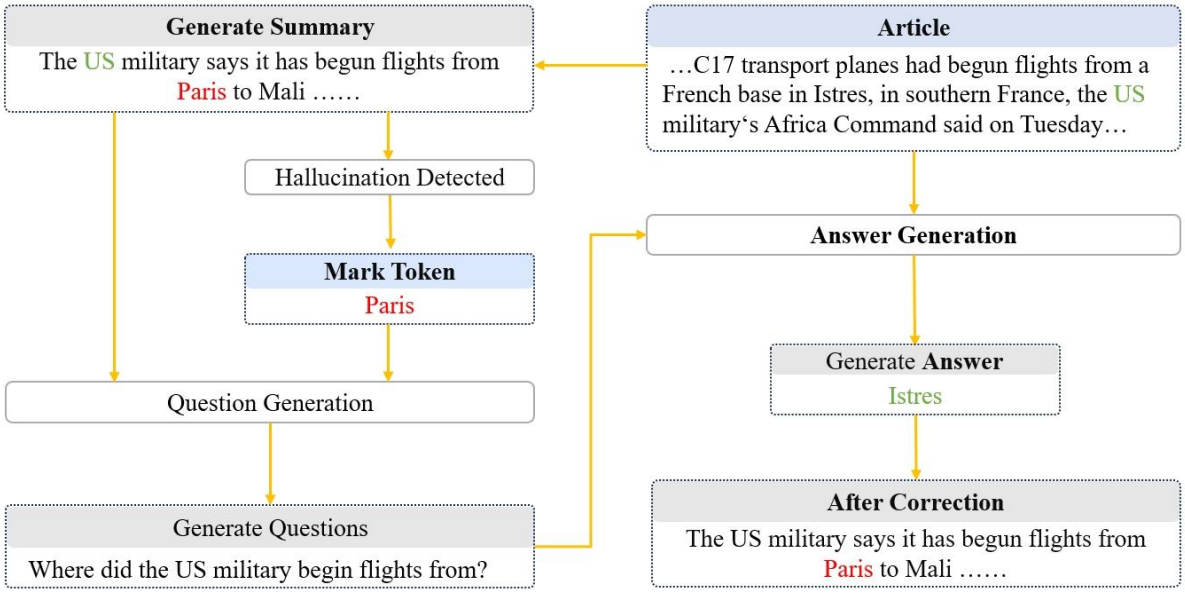

Figure (1): Model Architecture Diagram

The figure demonstrates a text summary correction example based on the XSum dataset, representing a typical external hallucination issue. In the generated summary, "The US military says it has begun flights from Paris to Mali...," the entity "Paris" is identified as a potential hallucination and flagged. The BART baseline model mistakenly labels the departure location as "Paris" (which does not appear in the original text). This may be due to the appearance of the related term "France" in the article, which influenced the BART model's generation process.

Next, the question generation module generates a targeted question: "Where did the US military begin flights from?" The question-answering model then extracts the correct answer from the original text: "C17

transport planes had begun flights from a French base in Istres, in southern France...” Finally, the summary is corrected to: “The US military says it has begun flights from Istres to Mali...”. This approach successfully corrects the hallucinated vocabulary while keeping the summary quality and sentence structure largely unchanged.

3.1 HALLUCINATED LEXICON DETECTION STAGE

The core task of this phase is to quantify the uncertainty in the text generation process of the generative model, in order to identify potential hallucinated content in the summary (i.e., information that is inconsistent with the original text facts), providing clear intervention targets for subsequent content correction. By analyzing the model's confidence level during generation, we can effectively locate words or segments with a higher risk of hallucinations, and then make targeted corrections in subsequent steps to improve the factual accuracy and overall reliability of the summary.

The model's uncertainty can be mainly categorized into two types: Aleatoric Uncertainty and Epistemic Uncertainty. The former is caused by noise or variations introduced during data collection or preprocessing, while the latter reflects knowledge gaps resulting from insufficient training or lack of experience.

In generative summarization tasks, the occurrence of hallucinations is often closely related to the aforementioned uncertainties. When the model lacks full confidence in its generated results, it tends to select high marginal probability words that frequently appear in the training corpus but may not be consistent with the context, which could lead to the generation of false information. The marginal probability here refers to the likelihood of the model generating a particular word at a given time step without considering the input text. The higher the uncertainty, the more likely this probability distribution deviates from the real context, leading to hallucinated content. Therefore, accurately assessing the uncertainty during the generation process helps to evaluate the risk of hallucinations.

To measure this uncertainty, we introduce Entropy as a unified evaluation metric. Entropy reflects both the Aleatoric and Epistemic uncertainties of the model, making it a robust generalization tool. In this study, we use entropy to identify potential hallucinated entities: when the entropy value at a particular generation step is high, it means the model's output at that step has greater uncertainty, thus the risk of hallucination is also higher; conversely, a low entropy value indicates a more confident output from the model. Based on this observation, we set an entropy threshold and label words with entropy values above this threshold as potential hallucinated content.

In the experimental section, we calculate the output information entropy of each time step during the generation process based on the Bart-Base model, as shown in Equation (1.1).

$$H(P(\square\, y < t, x)) = -\sum_{y \in v} p(y \mid y_{<t}, x) \log p(y \mid y_{<t}, x) \qquad (1.1)$$

Here, x represents the source text, and $p(y \mid y_{<t}, x)$ indicates the probability of the model generating a word $H(P(\square\, y < t, x))$ at a specific time step. The overall uncertainty at the current step is computed by aggregating the entropy over all candidate words at that time step.

Since hallucinated content often arises from incorrect combinations of entities (i.e., internal hallucinations) or the generation of unsupported entities (i.e., external hallucinations), it is particularly important to perform factual correction with entities as the core unit. When certain subwords within an entity exhibit high information entropy during generation, it indicates that the model experiences significant uncertainty at those time steps, leading to a higher average entropy for the entity as a whole. This suggests that the model lacks stability and confidence when generating that entity. Therefore, in this study, we treat the entity as the fundamental unit and aggregate the information entropy of its constituent subwords to identify potentially hallucinated terms. Let entity E consist of *n* subwords, and its average entropy is calculated as shown in Equation (1.2).

Additionally, we observed that hallucinations are more likely to occur in the latter half of the summaries in the XSum dataset. Therefore, we specifically focus on the entropy distribution in this section and mark the largest entity whose information entropy exceeds a given threshold to further analyze its potential hallucination risk. Let entity E consist of n subwords, and its average entropy is calculated as shown in Equation (1.2).

$$H(E) = \frac{1}{n} \sum_{i=1}^{n} H(x_i) \qquad (1.2)$$

To accurately identify key entities, we incorporate the NLTK toolkit for Named Entity Recognition (NER), and only annotate the named entities detected in the summary. This approach helps reduce noise and enhances the model’s focus on critical content. By doing so, the model gains a better understanding of the factual alignment between the source text and the summary, thereby improving the effectiveness of hallucination detection and correction.

We adopt a conservative tagging strategy: as shown in Equation (1.3), only the first entity with the highest uncertainty is marked among those whose average entropy exceeds a predefined threshold. This strategy helps prevent overcorrection and ensures that the model concentrates on the most error-prone segments, ultimately improving both the accuracy and efficiency of the correction process.

$$Mark(E) = \begin{cases} 1 & if\ H_E > \theta\ and\ H_E = \max(H_{E_1}, ..., H_{E_K}) \\ 0 & otherwise \end{cases} \quad (1.3)$$

3.2 CORRECTION PHASE FOR HALLUCINATED EXPRESSIONS

Inspired by the human reasoning process in identifying and correcting factual errors, this model introduces a "question generation–answer generation" approach to address potential hallucinations in generated summaries. Centered around the detected hallucinated entities, the method constructs targeted questions to guide the model in understanding the context and producing answers that align with the source text. These answers are then used to replace the hallucinated content, thereby correcting factual inconsistencies while preserving the original sentence structure as much as possible. This approach enhances the factual consistency and reliability of the summary. As illustrated in the figure, the entire process consists of two main stages:

(1) QUESTION GENERATION STAGE

We use the original text as the basis for querying, with the goal of extracting or verifying factual information. The questions are crafted to align semantically with the declarative answers, while fully leveraging the contextual clues in the original statements. In this process, a question generation model is used. For each declarative answer extracted from the text, we combine it with its corresponding input statement to serve as input to the question generator, which then produces a corresponding question.

This approach ensures that the generated questions are not only tightly focused on specific answer content but also remain consistent with the original context. The goal is to produce questions that are precise enough to retrieve or validate the correct information units from the source text.

To handle various types of factual inconsistencies and answer formats, we adopt MixQG[26] (Murakhovs'ka et al., 2022) as our question generation model G. MixQG is trained on a mixture of nine different question-answering datasets, covering a wide range of answer types including boolean, multiple-choice, extractive, and abstractive formats. This diverse training enables MixQG to effectively address various factual error patterns and generate contextually appropriate and semantically clear questions.

(2) ANSWER GENERATION STAGE

During the answer generation phase, the experiment performs reasoning based on the previously generated questions and the original text, identifying and extracting accurate information units to replace hallucinated entities in the summary. This process relies on a question answering (QA) model to extract factual answers corresponding to each question from the given original text E. The questions are concatenated with the evidence and then fed into the QA model to obtain the corresponding answers.

For model selection, we adopted a QA model based on the T5 architecture. Specifically, in some experiments, we used UnifiedQA-v2[27] (Khashabi et al., 2022), a general-purpose QA model built on T5. It has been jointly trained on twenty different QA datasets, enabling it to handle a wide range of question types and making it suitable for various application scenarios.

IV. EXPERIMENT

To validate the effectiveness of the proposed method, this section presents experimental analyses from three perspectives. First, we investigate the relationship between the frequency of hallucinations and information entropy across different datasets, aiming to reveal the role of entropy in the hallucination generation mechanism. Second, we evaluate the impact of varying entropy thresholds on hallucination detection accuracy, in order to determine the optimal threshold setting. Finally, a comprehensive assessment of the proposed method's effectiveness in enhancing factual consistency is conducted through comparative experiments against baseline models and existing hallucination correction approaches, focusing on both summary quality and factual accuracy.

4.1 DATA SETS AND DIVISION OF DATA SETS

In this study, experiments were conducted using the publicly available XSum dataset for abstractive text summarization. The XSum dataset was constructed based on BBC news articles, where each article consists of a news body and a corresponding summary. Notably, the summary is typically the first sentence of the article and is referred to as a "single-sentence extreme summary." Compared to extractive summarization datasets, XSum

represents a more challenging task due to its abstractive nature, featuring shorter and more diverse summaries without fixed sentence structures. The dataset contains a total of 204,045 training samples, 11,332 validation samples, and 11,334 test samples, with average lengths of 411 tokens for articles and 23 tokens for summaries.

4.2 EXPERIMENTAL SETTINGS

We randomly initialized the experimental environment five times and reported the average results. The pretrained BART-large model was fine-tuned on the XSum text summarization dataset using the Huggingface toolkit. Model training was conducted on Two NVIDIA RTX 4090 GPU. During training, the learning rate was set to 3e−5, with 5 epochs, and a batch size of 8. During the generation phase, we set the beam search size to 5 and the maximum generation length to 256. These settings aim to balance the diversity and accuracy of the generated summaries.

4.3 EVALUATION INDICATORS

(1)ROUGE

ROUGE[27] is used to measure the textual overlap between generated summaries and reference summaries. It mainly includes three sub-metrics: ROUGE-1, ROUGE-2, and ROUGE-L, which reflect n-gram matches and longest common subsequence alignment.

(2)FactCC

FactCC [13] is a sentence-level evaluation method that uses BERT to assess the factual consistency of generated summaries. It is specifically designed to identify whether the generated summary maintains factual consistency with the source text. Unlike ROUGE, which mainly focuses on surface-level overlap, FactCC places more emphasis on the semantic-level factual accuracy of the generated content.

(3)FactScore

FactScore[28] is a word- or phrase-level factual consistency metric based on pretrained language models. It can accurately locate specific factual errors within summaries and provides a more fine-grained assessment of the factual reliability of generated content.

4.4 COMPARE TO BASELINE

To evaluate the effectiveness of the proposed strategy, this section selects representative generative summarization models as baseline systems. The selected baselines include:

1. QA-SPAN and Autoregressive strategies proposed by Dong et al[14]., which are post-editing approaches designed to correct hallucinations. These methods iteratively identify all entities and use two separate pointer networks to predict the start and end positions of the answer spans. In our experiments, the base summarization model is replaced with BART-base, while the post-editing strategies follow the original implementation.
2. The Factual Corrector (FC) post-editing strategy proposed in the FASUM framework by Zhu et al[30]., which enhances factual consistency without modifying the original summarization architecture. It directly edits the candidate summaries generated by any abstractive summarization model. In our experiments, we adopt BART-base as the generation model and apply the FC module for post-editing as a comparison.

4.5 EXPERIMENTAL RESULTS

In this experiment, the BART-base model was trained on the XSum dataset to generate summaries on the test set, with the goal of analyzing the relationship between entropy and hallucinated content.

As illustrated in Figure 2, a stage-wise analysis of the generation process was conducted by dividing each generated summary into five equal-length segments, each representing 20% of the total output. For each segment, we computed the proportion of hallucinated tokens and their corresponding average entropy. The results reveal a clear trend: entropy gradually increases during the first 80% of the generation process, peaking between the 60–80% segment at 2.59. This segment also exhibits the highest hallucination rate at 37%. In the final 80–100% segment, both metrics decline, forming a rise-and-fall pattern. This suggests a strong positive correlation between entropy and hallucination—the higher the entropy, the more likely the model is to generate hallucinated content.

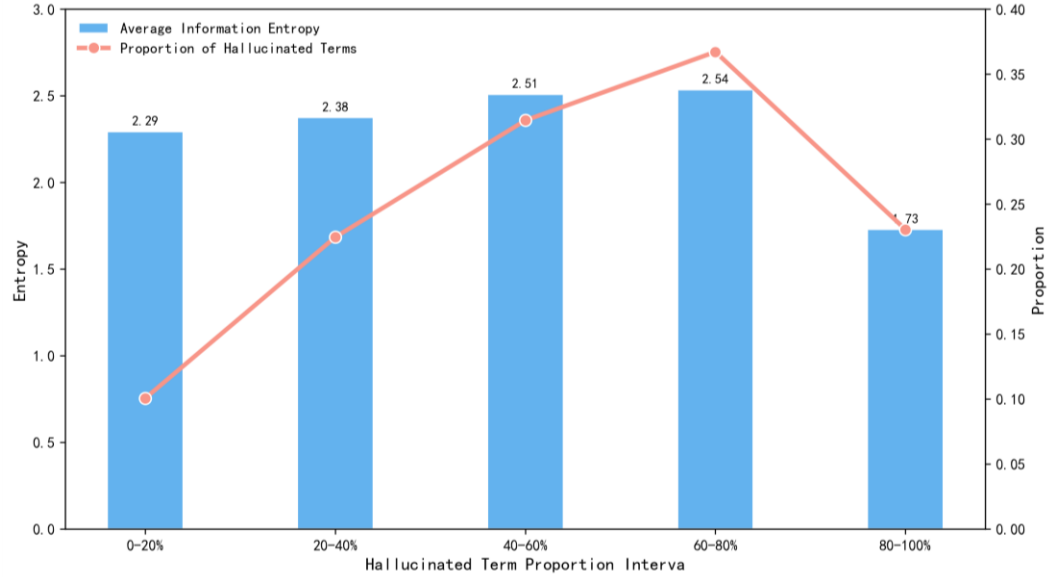

Figure (2): Distribution of Hallucinated Term Entropy and Proportion in the XSum Dataset

Complementing this trend, Table 1 presents the overall entropy statistics across the entire dataset. On average, hallucinated words exhibit a significantly higher entropy (2.43) compared to non-hallucinated words (2.10), while the overall average word entropy stands at 2.28. Notably, hallucinated content makes up approximately 32% of the total words in the generated summaries.

Table 1: Hallucination vs. Non-Hallucination Entropy Statistics

Evaluation Type	Entropy
Non-hallucinated Words	2.10
Hallucinated Words	2.43
Average Entropy	2.28
Hallucination Rate	32%

To further validate this observation, we set an entropy threshold to identify hallucinated entities and evaluated the precision of hallucination detection. On the XSum dataset, the initial threshold was set to 2.0, and incrementally increased by 0.1. We then measured the detection precision at each threshold level. Detailed results are presented in Figure 3.

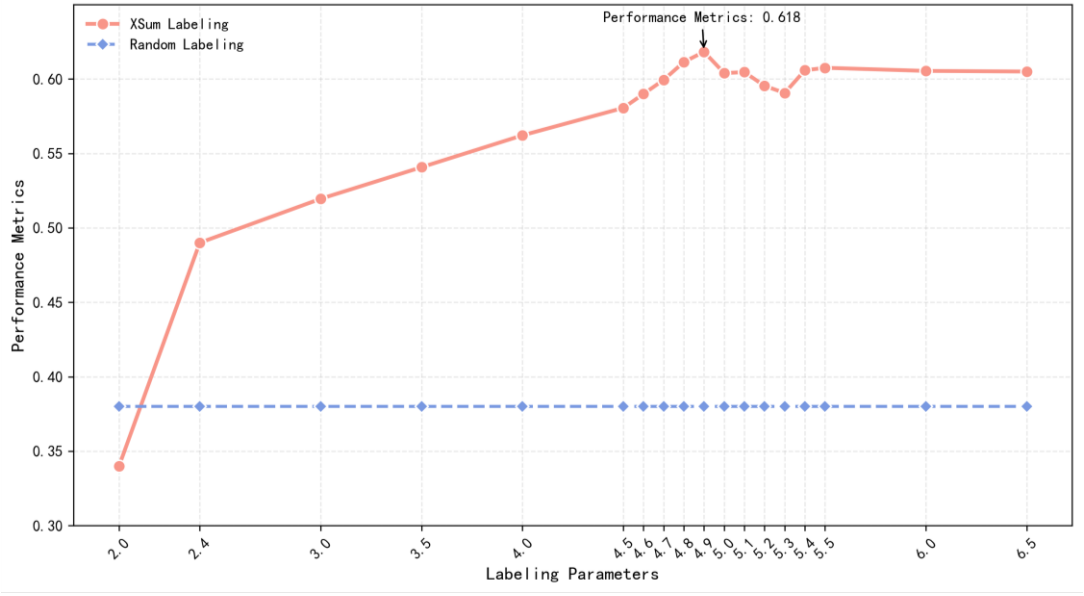

Figure (3): Performance Comparison of Different Labeling Methods

On the XSum dataset, the precision of randomly labeling entities is 0.38. In contrast, the precision of entropy-based labeling gradually improves as the entropy threshold increases and stabilizes around a value of 4.7. The highest precision is achieved at a threshold of 4.9, reaching 0.618. The final precision values under different entropy thresholds are presented in Table 2.

To assess the effectiveness of our proposed method, we conducted comparative experiments against baseline models. For factual consistency evaluation, we employed FactCC and FactScore as metrics, and for quality assessment, we utilized ROUGE metrics. The experiments were performed on the XSum dataset, with results presented in Table 3.6. In this table, R-1, R-2, and R-L correspond to ROUGE-1, ROUGE-2, and ROUGE-L scores, respectively, while FCC and FS denote FactCC and FactScore.

Table 2: Comparative Results of Different Models on XSum definitions

Model	Quality Score			Factuality Score	
	R-1	R-2	R-L	FCC	FS
XSum					
BART-Base	45.65	21.47	**37.14**	26.54	67.26
QA-SAPN	**45.82**	21.52	37.12	27.70	67.95
Autoregressive	45.40	**21.54**	37.01	26.93	67.95
FASUM	44.63	21.28	36.53	**27.91**	68.31
EQFCS	45.74	22.10	37.09	27.40	**70.83**

The results in the table indicate that the proposed method significantly improves factual consistency while maintaining the overall quality of the generated summaries. On the XSum task, our method achieves comparable performance to the baseline in terms of the ROUGE quality metric. However, it demonstrates clear advantages in factual consistency, with improvements of 0.86 and 3.57 over the baseline on the FactCC and FactScore metrics, respectively.

To evaluate the effectiveness of entropy-based tagging of potential hallucinated entities for downstream correction, we first conducted a control experiment by randomly masking entities in the generated summaries (Random) and applying the same question-answering model for factual correction. Additionally, to assess the impact of selectively masking only high-entropy entities, we introduced a second baseline (Iterate Mask), where all entities exceeding a predefined entropy threshold were iteratively masked and corrected. These were compared against our proposed method (EQFCS), and the results are shown in the table 3.

Table 3: Ablation Studys

Dataset	Factuality Score (FCC/FS)		
	Random	Iterate Mask	EQFCS
XSum	26.7/67.41	27.1/68.32	**27.4/70.83**

As shown in the table, the entropy-guided entity selection method (EQFCS) outperforms both baseline strategies in factual consistency across both evaluation metrics. These findings confirm the effectiveness of using information entropy to accurately identify and correct potential hallucinated entities.

4.6 CASE ANALYSIS

As shown in Table 4, This example demonstrates the effectiveness of the proposed method in correcting internal hallucinations. In the original output, the BART model incorrectly identified Chelsea manager Jose Mourinho as the manager of West Ham, which is a typical case of internal hallucination—confusing factual information already present in the text. By generating the question "Who is the manager of West Ham?" and obtaining the correct answer "Allardyce," the model was able to detect the role attribution error and revise the summary to "West Ham manager Allardyce says...," successfully correcting the factual mistake. This process highlights the effectiveness of the proposed method in detecting and correcting internal hallucinations.

Table 4: CASE ANALYSIS

Summary : ...I told Big Sam [West Ham manager Allardyce] and I repeat my words: they need points and, because they need points, to come here and play the way they did, is it acceptable? Maybe, yes. …
Reference Summary: Jose Mourinho accused West Ham of playing "19th-Century football", after his Chelsea side were held 0-0 in Wednesday's league encounter.
BART : West Ham manager Jose Mourinho says he was "outwitted" by manager Sam Allardyce after his side were held to a goalless draw by Chelsea.
Hallucination Tag : Jose Mourinho
Question Generation: Who is the manager of West Ham?

Answer Generation: Allardyce
After Correction: West Ham manager Allardyce says he was "outwitted" by manager Sam Allardyce after his side were held to a goalless draw by Chelsea.

V. CONCLUSION

This work introduces a novel approach that combines information entropy with a question-answering (QA) mechanism to improve factual consistency in text summarization. An entropy-based detection method is developed to effectively locate potential hallucinated entities in generated summaries. Building on this, a QA-driven correction framework is constructed, simulating human reasoning to automatically revise hallucinated content and replace it with accurate factual information. Experimental evaluations confirm that the method enhances factual accuracy while preserving the summary's grammatical structure and fluency.

Despite these improvements, addressing deeper semantic hallucinations remains a challenge—especially those beyond entity-level inconsistencies. The current QA-based correction mainly focuses on named entities, leaving issues like causal misinterpretation, temporal errors, and sentiment mismatches largely unaddressed. Future research will aim to refine the classification of hallucination types, such as subject-predicate mismatch, temporal disorder, and incorrect attribution. This will be supported by techniques like dependency parsing, event extraction, and semantic role labeling, allowing for more precise detection and targeted correction. Incorporating neural-symbolic reasoning and multi-step inference frameworks is also expected to enable more comprehensive modeling of complex hallucinations and enhance overall factual soundness.

Currently, the framework relies heavily on the source text as its factual reference, which can limit performance in scenarios requiring broader context or external knowledge. To address this, future efforts will focus on multi-source factual enhancement by integrating external knowledge bases (e.g., Wikidata), cross-document context, narrative chains, and domain-specific materials. Fusing these heterogeneous information sources will not only enrich semantic understanding but also fill potential factual gaps, thereby strengthening logical grounding and further improving summary reliability in complex environments.

REFERENCES

[1]. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27.
[2]. Rush A M, Chopra S, Weston J. A neural attention model for abstractive sentence summarization[J]. arXiv preprint arXiv:1509.00685, 2015.
[3]. Chen Q, Zhu X, Ling Z H, et al. Distraction-Based Neural Networks for Modeling Document[C]//IJCAI. 2016, 16: 2754-2760.
[4]. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[5]. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//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: 4171-4186.
[6]. Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. arXiv preprint arXiv:1910.13461, 2019.
[7]. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(140): 1-67.
[8]. Zhang J, Zhao Y, Saleh M, et al. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization[C]//International conference on machine learning. PMLR, 2020: 11328-11339.
[9]. Rohde T, Wu X, Liu Y. Hierarchical learning for generation with long source sequences[J]. arXiv preprint arXiv:2104.07545, 2021.
[10]. Dixit T, Wang F, Chen M. Improving factuality of abstractive summarization without sacrificing summary quality[J]. arXiv preprint arXiv:2305.14981, 2023.
[11]. Ryu S, Do H, Kim Y, et al. Multi-dimensional optimization for text summarization via reinforcement learning[J]. arXiv preprint arXiv:2406.00303, 2024.
[12]. Cao M, Dong Y, Wu J, et al. Factual error correction for abstractive summarization models[J]. arXiv preprint arXiv:2010.08712, 2020.
[13]. Kryściński W, McCann B, Xiong C, et al. Evaluating the factual consistency of abstractive text summarization[J]. arXiv preprint arXiv:1910.12840, 2019.
[14]. Dong Y, Wang S, Gan Z, et al. Multi-fact correction in abstractive text summarization[J]. arXiv preprint arXiv:2010.02443, 2020.
[15]. Fabbri A R, Choubey P K, Vig J, et al. Improving factual consistency in summarization with compression-based post-editing[J]. arXiv preprint arXiv:2211.06196, 2022.
[16]. Dong Y, Wieting J, Verga P. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization[J]. arXiv preprint arXiv:2204.13761, 2022.
[17]. Cao Z, Wei F, Li W, et al. Faithful to the original: Fact aware neural abstractive summarization[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
[18]. Gunel B, Zhu C, Zeng M, et al. Mind the facts: Knowledge-boosted coherent abstractive text summarization[J]. arXiv preprint arXiv:2006.15435, 2020.
[19]. Song J, Park N, Hwang B, et al. Model Intrinsic Features of Fine-tuning based Text Summarization Models for Factual Consistency[C]//Findings of the Association for Computational Linguistics: ACL 2023. 2023: 13884-13898.
[20]. Choubey P K, Fabbri A R, Vig J, et al. Cape: Contrastive parameter ensembling for reducing hallucination in abstractive summarization[J]. arXiv preprint arXiv:2110.07166, 2021.
[21]. Dixit T, Wang F, Chen M. Improving factuality of abstractive summarization without sacrificing summary quality[J]. arXiv preprint arXiv:2305.14981, 2023.

[22]. Liu Y, Deb B, Teruel M, et al. On improving summarization factual consistency from natural language feedback[J]. arXiv preprint arXiv:2212.09968, 2022.
[23]. Zha Y, Yang Y, Li R, et al. AlignScore: Evaluating factual consistency with a unified alignment function[J]. arXiv preprint arXiv:2305.16739, 2023.
[24]. Falke T, Ribeiro L F R, Utama P A, et al. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019: 2214-2220.
[25]. Roit P, Ferret J, Shani L, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback[J]. arXiv preprint arXiv:2306.00186, 2023.
[26]. Murakhovs' ka L, Wu C S, Laban P, et al. Mixqg: Neural question generation with mixed answer types[J]. arXiv preprint arXiv:2110.08175, 2021.
[27]. Khashabi D, Kordi Y, Hajishirzi H. Unifiedqa-v2: Stronger generalization via broader cross-format training[J]. arXiv preprint arXiv:2202.12359, 2022.
[28]. Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.
[29]. Zhou C, Neubig G, Gu J, et al. Detecting hallucinated content in conditional neural sequence generation[J]. arXiv preprint arXiv:2011.02593, 2020.
[30]. Zhu C, Hinthorn W, Xu R, et al. Enhancing factual consistency of abstractive summarization[J]. arXiv preprint arXiv:2003.08612, 2020.