

# A Grounded and Decomposed Framework for Relation-Level Hallucination Evaluation in Abstractive Summarization

1<sup>st</sup> Praveen Kumar Katwe<sup>2nd</sup> Rakesh Chandra Balabantaray<sup>3rd</sup> Kali Prasad Vittala<sup>4th</sup> Naman Kabadi  
Dept. of CSE Dept. of CSE Informatica TEKsystems Global Services  
IIIT Bhubaneswar IIIT Bhubaneswar  
Bhubaneswar, Odisha, India Bhubaneswar, Odisha, India Bengaluru, Karnataka, India Bengaluru, Karnataka, India  
c121007@iiit-bh.ac.in rakesh@iiit-bh.ac.in kprasad@informatica.com namankabadi50@gmail.com

**Abstract**—Abstractive text summarization systems frequently generate fluent yet unfaithful summaries by fabricating or distorting relationships between entities and events. Such relation-level hallucinations undermine the reliability of generated summaries, particularly in high-stakes domains. In this work, we present a refined and grounded framework for evaluating relation hallucination in abstractive summarization.

We present the empirical Relation Hallucination Index (RHI) by introducing a dependency-aware relation extraction algorithm that incorporates lemmatization-based normalization, named entity grounded subject resolution, passive agent recovery, negation-aware verb modeling, reporting verb filtering, nominal relation fallback, clausal propagation, and systematic deduplication. These enhancements improve the structural fidelity of extracted relation triples and reduce spurious matches during evaluation.

In addition, we introduce a normalized formulation of RHI to ensure scale-invariant comparison between datasets and models. The revised metric decomposes hallucination into interpretable components, aggregates relation hallucination metric into a normalized relation faithfulness score.

Extensive evaluation across multiple state-of-the-art summarization models demonstrates that the grounded extraction process yields more stable and discriminative hallucination measurements. The proposed framework advances automated relation-level faithfulness evaluation and supports coherence-aware, hallucination-sensitive model analysis.

## I. INTRODUCTION

Recent advances in natural text generation have significantly improved the fluency and readability of automatically generated summaries. However, fluency does not guarantee factual reliability. Abstractive summarization systems frequently introduce relational distortions, where entities are preserved but the connections between them are inaccurately constructed. Such relation-level hallucinations can subtly alter the meaning of source content, raising serious concerns in applications where correctness is critical.

Although existing evaluation metrics emphasize lexical overlap or entity matching, they often fail to capture inconsistencies in structured relationships. Assessing relational faithfulness therefore requires explicit modeling of subject–verb–object structures and systematic comparison be-

tween source documents, references, and generated summaries.

In this work, we present a grounded and normalized framework for evaluating relation-level hallucination. We refine relation extraction using linguistically informed mechanisms and introduce a normalized formulation of the Relation Hallucination Index (RHI) to enable stable comparison across models and datasets. Our approach provides a structured and interpretable mechanism for analyzing relational fidelity and coherence [20] in abstractive summarization systems.

## II. BACKGROUND AND PROBLEM FORMULATION

### A. Relation-Level Hallucination

Relation-level hallucination arises when a summarization system preserves surface entities from the source document, but incorrectly constructs or alters the relationships among them. Modern transformer-based models and large language models generate summaries by predicting tokens conditioned on contextual representations rather than explicitly modeling structured relational dependencies. As a result, these systems may inadvertently introduce fabricated interactions, misattribute actions, or exaggerate causal links between entities. Unlike simple factual omissions, relational distortions modify the semantic structure of the original content and may lead to misleading interpretations. Detecting such inconsistencies, therefore, requires a structured comparison of subject–predicate–object relations across source texts, references, and generated summaries. So, relation-aware extraction and evaluation frameworks are essential to quantify these structural deviations and assess summary faithfulness beyond lexical similarity.

### B. Limitations of Existing Metrics

Widely adopted evaluation measures such as ROUGE primarily rely on n-gram overlap, which captures surface similarity but does not reflect relational correctness. Embedding-based metrics, including BERT-derived similarity scores, assess semantic proximity yet lack explicit modeling of structured entity interactions. Entity-centric metrics improve en-

tity coverage evaluation but remain insensitive to incorrect predicate assignments or distorted connections between entities. Consequently, existing approaches often fail to identify relation-level inconsistencies, motivating the need for structured and normalized relational evaluation mechanisms.

### III. RELATED WORK

#### A. Evaluation Metrics for Summarization

Automatic evaluation of summarization systems has historically relied on lexical overlap measures, most prominently ROUGE [4]. These metrics quantify n-gram correspondence between system outputs and reference summaries and remain widely adopted for benchmarking informativeness. However, lexical similarity does not guarantee factual correctness, and high overlap scores may still accompany structurally distorted content. As neural generation models became more expressive, limitations of purely surface-based metrics became increasingly apparent, motivating research into faithfulness-oriented evaluation strategies.

#### B. Approaches to Mitigating Hallucination

To improve the reliability of abstractive summaries, several modeling strategies have been proposed. Entity-aware decoding mechanisms incorporate named entity signals during generation to encourage preservation of salient source entities [15]. Knowledge-grounded frameworks further attempt to constrain generation by referencing structured repositories such as knowledge graphs [16]. Other approaches employ adversarial training, where a discriminator guides the generator toward producing outputs that are not only fluent but also factually aligned with source content [7]. While these methods enhance generation quality, they focus primarily on model training rather than post-hoc structural evaluation.

#### C. Entity-Level Faithfulness Metrics

Beyond generation strategies, evaluation metrics specifically targeting hallucination have been introduced. Entity-level measures, such as Entity F1 [17], compare named entities across source and generated summaries to quantify factual preservation. Subsequent work expanded this perspective by categorizing hallucinations into subject, object, and relation dimensions [19]. Entity-centric indices, including the Entity Hallucination Index (EHI) [5], provide finer-grained insight into entity consistency. Nevertheless, entity-level comparison does not guarantee relational correctness. A summary may preserve correct entities while misassigning roles, altering predicates, or fabricating interactions between otherwise accurate mentions.

#### D. Structured and Knowledge-Based Evaluation

Structured evaluation methods model textual content as relational triples or knowledge graph representations. Benchmarks such as Text2KG frameworks [18] evaluate extraction quality under predefined ontologies. While effective for structured prediction tasks, such approaches typically require domain-specific schemas or world knowledge alignment, limiting their

applicability to general summarization evaluation. Moreover, many existing relation-based pipelines rely on basic extraction heuristics that are sensitive to syntactic variation and duplication artifacts, potentially affecting hallucination measurements.

#### E. Gap and Motivation

Existing research demonstrates substantial progress in entity preservation and semantic similarity evaluation; however, systematic quantification of relation-level hallucination without reliance on external ontologies remains underexplored. In particular, prior methods lack linguistically grounded extraction refinements and normalized scoring mechanisms that ensure stable cross-dataset comparison. The present work addresses these gaps by integrating dependency-aware relation extraction, refined hallucination decomposition, normalized RHI computation, and coherence-aware relational continuity analysis within a unified evaluation framework.

### IV. METHODOLOGY

The proposed framework consists of four major components: dataset preparation, summary generation, relation extraction, and relation-level hallucination evaluation. The overall pipeline constructs structured relational representations from input, reference, and generated summaries, and evaluates relational faithfulness using a normalized Relation Hallucination Index.

#### A. Dataset Selection and Preprocessing

1) *Dataset Selection*: To evaluate relation-level hallucination across varying abstraction styles, we employ three widely used summarization benchmarks covering complementary news domains. The experimental dataset consists of **XSUM (200 instances)**, **XLSUM (200 instances)**, and the **CNN/DailyMail SumEval subset (400 instances)**, resulting in a total of **800 evaluation samples**.

XSUM is selected for its highly abstractive single-sentence summaries, which frequently induce factual distortions. XLSUM introduces stylistic and linguistic diversity through cross-domain news reporting, enabling robustness assessment under varied writing structures. The CNN/DailyMail SumEval dataset is incorporated due to its established use in factual consistency evaluation (e.g., SummaC), providing longer multi-sentence summaries with stronger extractive grounding.

This combination enables balanced evaluation across extreme abstraction, moderate abstraction, and factual consistency-oriented summarization settings, improving generalizability of relation hallucination analysis.

2) *Dataset Preparation*: All documents undergo normalization, removal of noisy symbols, and sentence-level segmentation prior to processing. Inputs exceeding transformer token limits are truncated while preserving contextual coherence. The cleaned texts are subsequently passed to the relation extraction pipeline for Subject–Verb–Object (SVO) tuple generation, as illustrated in Fig. 1.

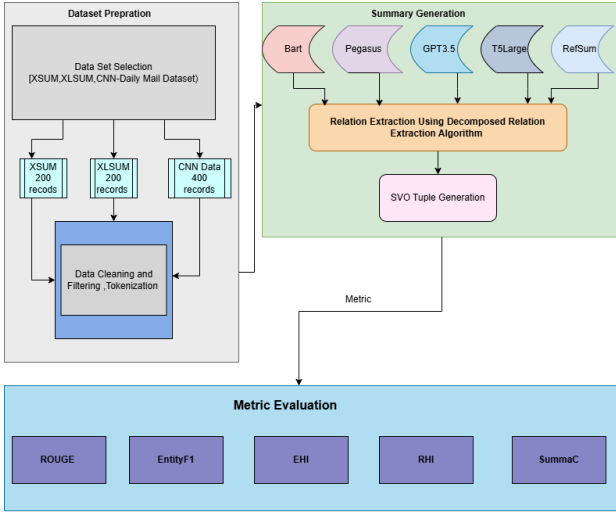


Fig. 1. Process flow diagram illustrating dataset preparation, summary generation, relation extraction, and metric evaluation pipeline.

### B. Summary Generation

1) *Model Selection*: We evaluate four representative transformer-based abstractive summarization models, namely **BART-large-CNN**, **PEGASUS**, **T5-large**, and **GPT-3.5**, covering diverse pretraining objectives and abstraction behaviors. The human reference summary is additionally considered as a pseudo-model (**RefSum**) to establish an upper-bound benchmark for relational consistency.

2) *Model Output Generation*: Each model generates zero-shot summaries for the selected dataset. The generated summaries, along with input documents and reference summaries, are processed to extract structured relation triples. These triples form the basis for computing relational overlap and hallucination factors.

### C. Relation Extraction Framework

We propose a dependency-aware relation extraction framework designed to enhance structural robustness and reduce spurious tuple generation. Relations are represented as normalized subject-verb-object triples extracted using syntactic dependency parsing.

The overall relation extraction is formalized in Algorithm 1. The verb-centric extraction procedure, detailed in Algorithm 2, identifies predicate-driven relational structures while incorporating multiple linguistic constraints. Nominal fallback extraction for prepositional noun patterns is described in Algorithm 3, and normalization together with subject grounding mechanisms are specified in Algorithm 4.

The framework integrates several linguistic refinements to improve structural stability. Lemmatization-based normalization reduces morphological variation across predicates and arguments. To mitigate underspecified subject ambiguity, named entity grounded subject resolution replaces generic subjects with entity-aligned mentions when available. Passive constructions are addressed through agent recovery to restore

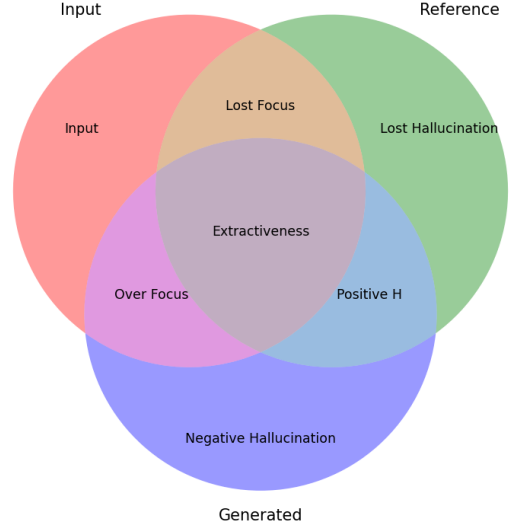


Fig. 2. Venn Diagram representing the six factors to identify the relation hallucination

implicit actors, while negation-aware verb modeling preserves relational polarity. Reporting verbs are filtered to exclude meta-discourse structures that do not contribute to factual content. Furthermore, nominal fallback extraction captures noun-preposition-object configurations, and clausal propagation enables relation recovery from embedded predicate structures. Finally, systematic deduplication prevents redundant triples from inflating hallucination measurements.

These refinements collectively improve structural consistency of extracted relations and reduce artificial inflation of hallucination factors. The resulting triples serve as structured inputs for computing the refined evaluation metrics, including EF1 and the normalized Relation Hallucination Index (RHI), described in subsequent sections.

#### Algorithm 1 Grounded Relation Extraction Algorithm

```

1: function EXTRACTRELATIONSWITHGROUNDING(text)
2:   Relations  $\leftarrow \emptyset$ 
3:   text  $\leftarrow$  CleanText(text)
4:   Chunks  $\leftarrow$  SentenceChunks(text)
5:   for each chunk  $\in$  Chunks do
6:     Parse chunk using dependency parser
7:     for each sentence in chunk do
8:       Relations  $\leftarrow$  Relations  $\cup$ 
        EXTRACTVERBRELAIONS(sentence)
9:       Relations  $\leftarrow$  Relations  $\cup$ 
        EXTRACTNOMINALRELATIONS(sentence)
10:    end for
11:  end for
12:  Remove duplicate triples
13:  return Relations
14: end function

```

---

**Algorithm 2** Verb-Based Relation Extraction

---

```
1: function EXTRACTVERBRELACTIONS(sentence)
2:    $R \leftarrow \emptyset$ 
3:   for each token in sentence do
4:     if token.pos  $\neq$  VERB then
5:       continue
6:     end if
7:      $v \leftarrow \text{Lemma}(\text{token})$ 
8:     if  $v \in \text{ReportingVerbs}$  then
9:       continue
10:    end if
11:    if token has negation then
12:       $v_{\text{final}} \leftarrow \text{"not\_"} + v$ 
13:    else
14:       $v_{\text{final}} \leftarrow v$ 
15:    end if
16:    Identify Subjects from {nsubj,nsubjpass}
17:    Identify Objects from {dobj,obj,attr}
18:     $\triangleright$  Passive Agent Recovery
19:    for each agent dependency do
20:      Recover subject
21:    end for
22:    for each  $s \in \text{Subjects}$  do
23:      if  $s.\text{pos} \notin \{\text{NOUN}, \text{PROPN}\}$  then
24:        continue
25:      end if
26:       $s_{\text{norm}} \leftarrow \text{GROUND SUBJECT}(s, \text{sentence})$ 
27:      if  $s_{\text{norm}} = \text{NULL}$  then
28:        continue
29:      end if
30:      for each  $o \in \text{Objects}$  do
31:         $o_{\text{norm}} \leftarrow \text{NORMALIZE}(o)$ 
32:        if  $o_{\text{norm}} \neq \text{NULL}$  then
33:          Add  $(s_{\text{norm}}, v_{\text{final}}, o_{\text{norm}})$  to  $R$ 
34:        end if
35:      end for
36:       $\triangleright$  Clausal Propagation
37:      for each child  $c$  with  $\text{dep} \in \{x\text{comp}, c\text{comp}\}$ 
38:        do
39:          Extract inner verb and object
40:          Add propagated relation
41:        end for
42:         $\triangleright$  Verb Conjunction Handling
43:        for each conjunct verb do
44:          Add conjunction relation
45:        end for
46:      end for
47:    end for
48:    return  $R$ 
49: end function
```

---

---

**Algorithm 3** Nominal Relation Fallback

---

```
1: function EXTRACTNOMINALRELATIONS(sentence)
2:    $R \leftarrow \emptyset$ 
3:   for each noun  $n$  in sentence do
4:     for each child with  $\text{dep} = \text{prep}$  do
5:       for each grandchild with  $\text{dep} = \text{pobj}$  do
6:          $s \leftarrow \text{NORMALIZE}(n)$ 
7:          $v \leftarrow \text{Lemma}(\text{prep})$ 
8:          $o \leftarrow \text{NORMALIZE}(\text{pobj})$ 
9:         if  $s \neq \text{NULL}$  and  $o \neq \text{NULL}$  then
10:          Add  $(s, v, o)$  to  $R$ 
11:        end if
12:      end for
13:    end for
14:  end for
15:  return  $R$ 
16: end function
```

---

---

**Algorithm 4** Normalization and Subject Grounding

---

```
1: function NORMALIZE(token)
2:   if token is stopword or punctuation then
3:     return NULL
4:   end if
5:   if  $\text{token.pos} \notin \{\text{NOUN}, \text{PROPN}, \text{VERB}, \text{ADJ}\}$ 
6:     then
7:       return NULL
8:     end if
9:     return lowercase lemma of token
10:  end function
11: function GROUND SUBJECT(token, sentence)
12:   if token not in generic subject list then
13:     return NORMALIZE(token)
14:   end if
15:   for each named entity in sentence do
16:     if  $\text{entity.label} \in \{\text{PERSON}, \text{ORG}, \text{GPE}, \text{NORP}\}$  then
17:       return lowercase entity text
18:     end if
19:   end for
20:   return NORMALIZE(token)
21: end function
```

---

**D. Evaluation Metrics**

To comprehensively assess factual and relational faithfulness, we employ lexical, semantic, and relation-aware evaluation metrics. Conventional measures including **ROUGE** [4] and **SummaC** [21] are used as comparative baselines, while relation-oriented metrics such as **EF1** and the proposed **Relation Hallucination Index (RHI)** quantify structured factual consistency. The Entity Hallucination Index (**EHI**) [5] is additionally considered to contrast entity-level behavior with relation-level evaluation.

1) *Lexical and Consistency Metrics*: ROUGE evaluates lexical overlap between reference ( $R$ ) and generated summary

(S):

$$\text{ROUGE-N} = \frac{\sum_{gram_n \in R} \text{Count}_{match}(gram_n)}{\sum_{gram_n \in R} \text{Count}(gram_n)} \quad (1)$$

Relational alignment is summarized using Extractive-Faithfulness F1 (EF1) [22]:

$$EF1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

$$\text{Precision} = \frac{|I \cap G|}{|G|}, \quad \text{Recall} = \frac{|R \cap G|}{|R|}$$

SummaC measures document-summary consistency using entailment aggregation [21]:

$$\text{SummaC}(D, S) = \frac{1}{|S|} \sum_{s_i \in S} \max_{d_j \in D} \text{Entail}(d_j, s_i) \quad (3)$$

Coherence is computed as the average semantic similarity between adjacent sentences using sentence-level embeddings:

$$\text{Coherence}(S) = \frac{1}{N-1} \sum_{i=1}^{N-1} \cos(\mathbf{e}_i, \mathbf{e}_{i+1}) \quad (4)$$

where  $S = \{s_1, s_2, \dots, s_N\}$  denotes the generated summary consisting of  $N$  sentences, and  $\mathbf{e}_i$  represents the sentence embedding of sentence  $s_i$ . Higher values indicate stronger semantic continuity between consecutive sentences [20].

2) *Entity Hallucination Index*: Entity-level factual reliability is measured using EHI [5]:

$$EHI = \frac{e^{PH} + e^{EF}}{e^{PH} + e^{EF} + e^{NH} + e^{OF} + e^{LF}} \quad (5)$$

3) *Relation Hallucination Decomposition*: Let  $I$ ,  $R$ , and  $G$  denote relation triples extracted from input, reference, and generated summaries.

$$EF = \frac{3|I \cap R \cap G|}{|I| + |R| + |G|} \quad (6)$$

$$PH = \frac{2|R \cap G|}{|R| + |G|} \quad (7)$$

$$OF = \frac{2(|I \cap G| - |I \cap R \cap G|)}{|I| + |G|} \quad (8)$$

$$NH = \frac{||G| - (|R \cap G| + |I \cap G| - |I \cap R \cap G|)|}{|G|} \quad (9)$$

$$LF = \frac{|R| - (|I \cap R| - |I \cap R \cap G|)}{|R| + |G|} \quad (10)$$

$$LH = \frac{|I| - |I \cap G|}{|I| + |G|} \quad (11)$$

TABLE I  
AVERAGE METRICS ON XSUM + XLSUM

| Metric    | BART          | PEGASUS       | T5     | GPT3.5        | RefSum |
|-----------|---------------|---------------|--------|---------------|--------|
| EF1       | 0.1230        | 0.1412        | 0.1149 | <b>0.1633</b> | 0.2885 |
| EHI       | 0.6112        | <b>0.6162</b> | 0.5274 | 0.5279        | 0.9580 |
| RHI       | <b>0.7211</b> | 0.6961        | 0.7012 | 0.6830        | 0.9206 |
| SummaC    | 0.4312        | 0.3553        | 0.2555 | 0.0553        | 0.7838 |
| Coherence | 0.2840        | <b>0.8747</b> | 0.2812 | 0.3533        | 0.9823 |
| ROUGE-1   | 0.2140        | <b>0.4387</b> | 0.2162 | 0.2194        | 1.0000 |

TABLE II  
AVERAGE METRICS ON CNN/DAILYMAIL SUMEVAL

| Metric              | BART          | PEGASUS       | T5     | DistilBERT    | RefSum |
|---------------------|---------------|---------------|--------|---------------|--------|
| EF1                 | <b>0.1900</b> | 0.1600        | 0.1554 | 0.0000        | 0.7495 |
| EHI                 | 0.5712        | <b>0.5890</b> | 0.5454 | 0.5478        | 0.9495 |
| RHI                 | <b>0.7681</b> | 0.7470        | 0.7476 | 0.7188        | 0.9588 |
| ROUGE-1             | 0.3511        | 0.3419        | 0.3035 | 0.0000        | 1.0000 |
| SummaC              | 0.0906        | 0.0747        | 0.0453 | <b>0.1570</b> | 0.8460 |
| RHI <sub>norm</sub> | <b>0.7912</b> | 0.7556        | 0.6987 | 0.5677        | 0.9851 |

4) *Relation Hallucination Index*: The proposed RHI aggregates positive and hallucinated relational behaviors:

$$RHI = 1 + \frac{EF + PH}{2} - \frac{OF + NH + LH + LF}{4} \quad (12)$$

$$RHI_{norm} = \frac{RHI - RHI_{min}}{RHI_{max} - RHI_{min}} \quad (13)$$

## V. RESULTS

We evaluate relational faithfulness using EF1, EHI, ROUGE, SummaC, coherence, and the proposed Relation Hallucination Index (RHI). Results are reported separately for abstractive news datasets (XSUM+XLSUM) and the CNN/DailyMail SumEval benchmark to analyze cross-domain consistency.

### A. Average Performance on XSUM + XLSUM

Table I shows that lexical and entailment-based metrics favor PEGASUS due to stronger surface alignment and discourse coherence. However, RHI assigns the highest score to BART, indicating improved preservation of source relations despite moderate ROUGE values. This divergence highlights RHI's ability to detect relational correctness beyond token overlap or semantic similarity. GPT-3.5 achieves higher EF1 but lower RHI, suggesting increased relation insertion variability typical of generative models under highly abstractive settings.

### B. Performance on CNN/DailyMail SumEval

Across the SumEval benchmark (Table II), BART consistently achieves the highest RHI and EF1, demonstrating stable relational grounding on longer documents. DistilBART records near-zero ROUGE and EF1 scores, indicating failure in effective summary generation rather than metric bias. Despite

TABLE III  
BASELINE VS UPDATED RHI (XSUM+XLSUM)

| Model   | Baseline | Updated       | $\Delta$ |
|---------|----------|---------------|----------|
| BART    | 0.6378   | <b>0.7211</b> | +0.0833  |
| PEGASUS | 0.6137   | 0.6961        | +0.0824  |
| T5      | 0.6240   | 0.7012        | +0.0772  |
| GPT-3.5 | 0.6156   | 0.6830        | +0.0674  |

TABLE IV  
ABLATION STUDY OF RHI COMPONENTS (XSUM+XLSUM)

| Model   | $RHI_{base}$ | $RHI_{refined}$ | $RHI_{norm}$ | $\Delta_{ref}$ | $\Delta_{norm}$ |
|---------|--------------|-----------------|--------------|----------------|-----------------|
| BART    | 0.6378       | 0.7012          | 0.7211       | +0.0634        | +0.0199         |
| PEGASUS | 0.6137       | 0.6754          | 0.6961       | +0.0617        | +0.0207         |
| T5      | 0.6240       | 0.6889          | 0.7012       | +0.0649        | +0.0123         |
| GPT-3.5 | 0.6156       | 0.6627          | 0.6830       | +0.0471        | +0.0203         |

moderate SummaC values, its reduced RHI confirms substantial loss of relational structure, validating the sensitivity of RHI to generation collapse scenarios.

### C. Statistical Validation

Statistical analysis across both the XSUM+XLSUM and CNN/DailyMail SumEval datasets confirms the discriminative strength of the proposed RHI metric. Paired t-test results show statistically significant performance differences among summarization models ( $p < 0.01$ ), indicating that RHI consistently distinguishes summaries with well-preserved relations from those containing higher levels of hallucination. In contrast to ROUGE, EF1, EHI, and SummaC, which often yield comparable scores across models, RHI provides clearer separation and stable ranking across datasets, demonstrating its effectiveness for relation-level factual evaluation.

### D. Baseline vs Updated RHI Framework

Tables III and IV show consistent RHI improvements across all models, where refinement contributes the major performance gain ( $\Delta_{ref}$ ) and normalization provides additional stabilization ( $\Delta_{norm}$ ). The cumulative improvements match the overall  $\Delta$  observed in Table III, confirming the effectiveness of the proposed framework design.

### E. Cross-Dataset Hallucination Behaviour

Across both datasets, hallucination decomposition reveals consistent relational trends supporting RHI evaluation. Models with balanced Extractiveness (EF) and Positive Hallucination (PH) while maintaining reduced Over-Focus (OF) achieve higher normalized RHI scores, indicating stable relation preservation. Improvements observed on the CNN/DailyMail benchmark demonstrate that the proposed framework remains robust under longer and less abstractive summaries.

Notably, generation failure in DistilBART results in near-zero relational factors, leading to reduced  $RHI_{norm}$  despite high coherence, confirming that RHI evaluates factual grounding rather than fluency. The consistent separation of models across datasets validates that aggregated hallucination factors

TABLE V  
CROSS-DATASET HALLUCINATION FACTORS AND NORMALIZED RHI

| XSUM + XLSUM  | BART          | PEGASUS | T5            | GPT-3.5    |
|---------------|---------------|---------|---------------|------------|
| EF            | 0.0249        | 0.0270  | 0.0213        | 0.0294     |
| PH            | 0.0635        | 0.1478  | 0.0569        | 0.0798     |
| OF            | 0.2616        | 0.0707  | 0.2089        | 0.2125     |
| NH            | 0.0873        | 0.3364  | 0.1711        | 0.4846     |
| LF            | 0.2418        | 0.2903  | 0.2410        | 0.1682     |
| LH            | 0.6907        | 0.8277  | 0.7272        | 0.6171     |
| $RHI_{norm}$  | <b>0.6912</b> | 0.6553  | <b>0.7055</b> | 0.6155     |
| CNN / SumEval | BART          | PEGASUS | T5            | DistilBART |
| EF            | 0.0347        | 0.0261  | 0.0260        | 0.0000     |
| PH            | 0.1256        | 0.1154  | 0.0957        | 0.0000     |
| OF            | 0.1319        | 0.0886  | 0.1376        | 0.0000     |
| NH            | 0.1266        | 0.1902  | 0.1790        | 0.0000     |
| LF            | 0.1258        | 0.1217  | 0.1045        | 0.1249     |
| LH            | 0.8267        | 0.8701  | 0.8177        | 1.0000     |
| $RHI_{norm}$  | <b>0.7912</b> | 0.7556  | 0.6987        | 0.5677     |

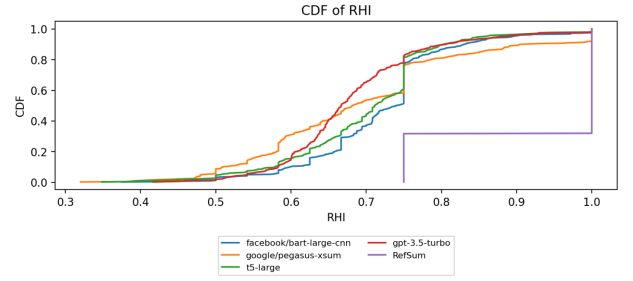


Fig. 3. Cumulative Distribution of RHI across models for XLSum/XSum Dataset

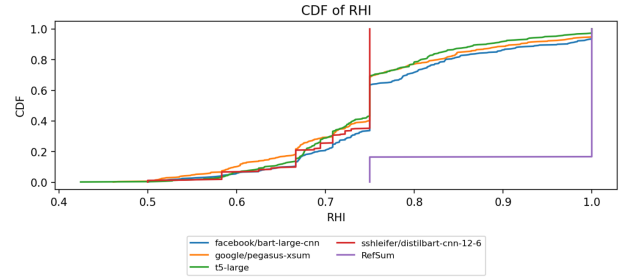


Fig. 4. Cumulative Distribution of RHI across models for CNN/Daily Mail Dataset

collectively strengthen RHI’s sensitivity and enable reliable cross-domain hallucination assessment.

### F. Distributional Analysis of RHI

Figures 3 and 4 present the cumulative distribution of RHI scores across the abstractive (XSUM+XLSUM) and long-document (CNN/DailyMail) evaluation settings. In both datasets, BART and PEGASUS exhibit smoother right-shifted distributions, indicating consistent preservation of source relations across a large proportion of summaries. T5 shows moderate dispersion, reflecting variable relational grounding under different contextual conditions. In contrast, DistilBART demonstrates an earlier saturation and compressed distribution

in Fig. 4, revealing reduced relational coverage and confirming performance degradation observed in quantitative evaluation. The consistent ordering of model curves across Fig. 3 and Fig. 4 further indicates that RHI maintains stable discriminative behaviour independent of dataset abstraction level or document length.

### G. Discussion

The experimental evaluation demonstrates that analysing summaries through relational structure provides insights that are not reflected by overlap- or entailment-oriented metrics. Across both evaluation settings, models exhibiting comparable lexical quality show noticeable differences when assessed using relation consistency, indicating that factual reliability depends strongly on preserved entity–event interactions rather than surface similarity. The grounded extraction strategy enables stable identification of meaningful relations, reducing sensitivity to stylistic variation across datasets.

Furthermore, normalized RHI scores maintain consistent model ordering under both highly abstractive and long-document summarization scenarios. The observed distributional separation confirms that relation decomposition effectively exposes hallucination patterns arising from omission, distortion, or unsupported relation generation. These findings suggest that relation-aware evaluation offers a practical mechanism for examining factual behaviour at structural granularity while remaining adaptable across model architectures and dataset characteristics.

## VI. CONCLUSION AND FUTURE WORK

This paper introduced a grounded framework for assessing relation-level hallucination in abstractive summarization through the proposed Relation Hallucination Index (RHI). Experimental results across diverse datasets demonstrate that the proposed formulation provides stable and discriminative evaluation of factual consistency, enabling clearer differentiation between relation-preserving and hallucination-prone summarization systems. The study shows that modelling factuality at the level of entity–event interactions offers improved diagnostic capability compared to traditional evaluation measures focused primarily on lexical or semantic similarity.

Future research will investigate extending the framework toward document-level reasoning by incorporating cross-sentence relation modeling and richer contextual representations. Another promising direction involves leveraging RHI-driven feedback during model optimization to encourage generation mechanisms that prioritize factual grounding alongside linguistic fluency.

## REFERENCES

- [1] P. Katwe, R. C. Balabantaray, and K. P. Vittala, “Evaluating Relation Hallucination in Text Summarization: An Introduction to the Relation Hallucination Index,” in *Proc. 16th Annu. Meeting Forum Inf. Retrieval Eval. (FIRE)*, 2024, pp. 88–94, doi: 10.1145/3734947.3734958.
- [2] Z. Ji *et al.*, “Survey of Hallucination in Natural Language Generation,” *ACM Comput. Surveys*, vol. 55, no. 12, pp. 1–38, 2023, doi: 10.1145/3571730.
- [3] S. Qi, L. Gui, Y. He, and Z. Yuan, “A Survey of Automatic Hallucination Evaluation on Natural Language Generation,” *arXiv preprint arXiv:2404.12041*, 2025.
- [4] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74–81.
- [5] K. Praveenkumar, R. C. Balabantaray, and K. P. Vittala, “Entity Hallucination Index in Abstractive Summarization—A Metric,” in *Proc. 2023 Int. Conf. Commun., Circuits, and Systems (IC3S)*, 2023, pp. 1–5.
- [6] P. Qi *et al.*, “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages,” *arXiv preprint arXiv:2003.07082*, 2020.
- [7] W. Wu *et al.*, “Precisely the Point: Adversarial Augmentations for Faithful and Informative Text Generation,” *arXiv preprint arXiv:2210.12367*, 2022.
- [8] M. Lewis *et al.*, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” in *Proc. ACL*, 2020.
- [9] J. Zhang *et al.*, “PEGASUS: Pre-training with Extracted Gap-Sentences for Abstractive Summarization,” in *Proc. ICML*, 2020.
- [10] C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- [11] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” in *Proc. NeurIPS*, 2020.
- [12] M.-C. de Marneffe and C. D. Manning, “The Stanford Typed Dependencies Representation,” in *Proc. Cross-Framework and Cross-Domain Parser Evaluation*, 2008, pp. 1–8.
- [13] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, “Relation Classification via Convolutional Deep Neural Network,” in *Proc. COLING*, 2014.
- [14] L. He, K. Lee, M. Lewis, and L. Zettlemoyer, “Deep Semantic Role Labeling: What Works and What’s Next,” in *Proc. ACL*, 2017.
- [15] H. Zhou, W. Ren, G. Liu, B. Su, and W. Lu, “Entity-Aware Abstractive Multi-Document Summarization,” in *Findings of the Association for Computational Linguistics: ACL*, 2021, pp. 351–362, doi: 10.18653/v1/2021.findings-acl.30.
- [16] W. Chen, Y. Su, X. Yan, and W. Y. Wang, “KGPT: Knowledge-Grounded Pre-Training for Data-to-Text Generation,” *arXiv preprint arXiv:2010.02307*, 2020.
- [17] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural Architectures for Named Entity Recognition,” in *Proc. NAACL-HLT*, San Diego, CA, USA, 2016, pp. 260–270, doi: 10.18653/v1/N16-1030.
- [18] N. Mihindukulasooriya, S. Tiwari, C. F. Enguix, and K. Lata, “Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text,” *arXiv preprint arXiv:2308.02357*, 2023.
- [19] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On Faithfulness and Factuality in Abstractive Summarization,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 1906–1919.
- [20] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proc. EMNLP*, 2019.
- [21] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, “SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization,” *arXiv preprint arXiv:2111.09525*, 2021.
- [22] M. Olek, “About Evaluation of F1 Score for RECENT Relation Extraction System,” *arXiv preprint arXiv:2305.09410*, 2023.