

CHI: A Composite Hallucination Index Unifying Entity, Relation, and Quantity Dimensions for Summarization Evaluation

Praveenkumar Katwe^{1,2}, Rakesh Chandra Balabantaray¹, Kali Prasad Vittala²

¹International Institute of Information Technology Bhubaneswar, India

²Salesforce India Pvt Ltd, Bengaluru, India

{c121007, rakesh}@iiit-bh.ac.in, {pkatwe, kvittala}@salesforce.com

Abstract

Faithfulness evaluation of abstractive summaries remains an open challenge, with existing metrics addressing only isolated hallucination types: factual entity errors, relational inconsistencies, or numerical fabrications, without capturing their co-occurrence or interaction. We introduce **CHI** (Composite Hallucination Index), the first unified hallucination metric that decomposes faithfulness errors into three orthogonal dimensions: entity hallucination (EHI), relation hallucination (RHI*), and quantity hallucination (QHI). Each dimension employs a shared softmax-normalized architecture over Venn diagram-derived factors representing extractiveness, positive hallucination, over-focus, negative hallucination, and lost focus. The novel QHI component introduces tolerance-aware numerical matching with exact, epsilon, derived, and temporal comparison modes. We fuse the three dimensions via harmonic mean to produce a single composite score that penalizes weakness in any dimension. We validate CHI on 800 source articles spanning four domains (news, medical, legal, financial) with summaries from five generation systems. Empirical results demonstrate that: (i) the three dimensions are statistically orthogonal ($\bar{\rho} = 0.148$), confirming they capture distinct error types; (ii) CHI achieves the highest system-level correlation with human judgments ($\rho = 0.66$, $p = 0.006$) on SummEval, outperforming ROUGE ($\rho = 0.53$), EHI ($\rho = 0.58$), and all individual components; and (iii) ablation studies confirm that all three dimensions contribute unique variance, with the full composite outperforming any individual component while providing decomposable error diagnostics unavailable from single-score baselines. CHI provides practitioners with a decomposable, interpretable, and efficient faithfulness metric suitable for both offline evaluation and online monitoring of summarization systems.

1 Introduction

Consider a financial summary that correctly identifies *Goldman Sachs* as the acquiring entity (no entity hallucination), accurately states the acquisition relationship (no relation hallucination), yet fabricates the deal value as \$4.2 billion when the source reports \$2.8 billion (quantity hallucination). Existing faithfulness metrics, designed to detect a single error type, would rate this summary as largely faithful, missing a critical numerical fabrication that could mislead investors. This example illustrates a fundamental limitation: hallucination is not monolithic but multi-dimensional, and evaluation must account for entity, relational, and numerical faithfulness simultaneously.

Large language models (LLMs) have dramatically improved the fluency and coherence of abstractive summarization, yet their propensity for hallucination remains a persistent challenge [14, 4]. A summary is unfaithful when it contains information unsupported by or contradicted by the source document, whether through fabricated entities, invented relationships, or incorrect quantities. The research community has responded with increasingly sophisticated faithfulness metrics, from overlap-based approaches like ROUGE [13] and BERTScore [23] to entailment-based methods such as SummaC [10] and AlignScore [22], and more recently to atomic-fact decomposition approaches including FActScore [15] and SAFE [21].

However, these metrics share a critical limitation: they treat hallucination as a unidimensional phenomenon. AlignScore measures overall alignment without distinguishing error types. FActScore decomposes claims into atomic facts but evaluates them against a single binary criterion. Even structured approaches like DAE [3] and SRLScore [2] focus on dependency or semantic-role representations without separately quantifying numerical accuracy. Recent meta-evaluation work by Kulkarni et al. [9] demonstrates that existing metrics are poorly correlated with each other (average pairwise $\rho < 0.45$), suggesting they capture different, but individually incomplete, facets of faithfulness.

The need for multi-dimensional evaluation is particularly acute in high-stakes domains. Medical summaries may fabricate drug dosages while correctly identifying the drug entity. Legal summaries may invert contractual relationships while preserving party names and monetary values. Financial summaries may introduce phantom entities in otherwise numerically accurate reports. Each error type carries distinct downstream risks, and a single aggregate score obscures actionable diagnostic information.

Contributions. We make three primary contributions:

1. **QHI (Quantity Hallucination Index):** We introduce the first dedicated metric for numerical hallucination in summarization, using tolerance-aware matching with exact, epsilon ($\pm 5\%$), derived (computationally related), and temporal comparison modes, instantiated within a softmax-normalized 5-factor architecture.
2. **Orthogonality Validation:** We provide empirical evidence that entity, relation, and quantity hallucination dimensions are statistically orthogonal ($\bar{\rho} < 0.3$, $p < 0.001$), confirming they capture genuinely distinct error phenomena rather than correlated signals.
3. **CHI (Composite Hallucination Index):** We unify EHI, RHI*, and QHI through a harmonic mean fusion that achieves the highest system-level correlation with human judgments ($\rho = 0.66$) among all evaluated metrics, while providing decomposable diagnostics at negligible computational cost.

The unified softmax architecture ensures all three dimensions are directly comparable; each produces a score in $[0, 1]$ with the same semantic interpretation: higher values indicate greater faithfulness (lower hallucination). This architectural consistency enables meaningful cross-dimensional analysis, ablation, and fusion without ad-hoc normalization.

2 Related Work

2.1 Faithfulness Metrics

Reference-free faithfulness evaluation has progressed from lexical overlap (ROUGE, BERTScore) through entailment classification to LLM-based verification. **AlignScore** [22] trains a unified alignment function over 4.7M examples from 15 NLI datasets, achieving strong correlation with human judgments at ACL 2023. **FActScore** [15] decomposes biographies into atomic facts verified against Wikipedia, establishing the decompose-then-verify paradigm at EMNLP 2023. **SAFE** [21] extends this with multi-step reasoning and web search verification, demonstrating superhuman accuracy on certain benchmarks at NeurIPS 2024. **MiniCheck** [19] distills GPT-4-level fact-checking into 7B parameter models for efficient deployment at EMNLP 2024.

While these metrics achieve impressive correlation with human judgments, they fundamentally produce a single scalar faithfulness score without decomposing errors by type. A summary with 3 entity errors and 0 numerical errors receives the same score as one with 0 entity errors and 3 numerical errors, despite these representing qualitatively different failure modes requiring different interventions.

2.2 Structured Evaluation Approaches

Structural representation approaches enable finer-grained analysis. **DAE** [3] at EMNLP 2020 annotates and classifies dependency arc entailment for each summary sentence. **SRLScore** [2] at *SEM 2023 decomposes faithfulness along semantic role dimensions (agent, patient, predicate). **AMRFact** [17] at NAACL 2024 leverages Abstract Meaning Representation for structured fact verification. Our own prior work introduced **EHI** [7] at ACI 2025 for entity-specific hallucination and **RHI** [6] at FIRE 2024 for relation-specific hallucination using linear normalization.

These methods demonstrate the value of typed decomposition but remain limited to a single structural dimension. CHI extends this line of work by integrating entity, relation, and quantity dimensions within a unified architecture.

2.3 Numerical Hallucination

Numerical faithfulness has received limited dedicated attention. The **NumEval** shared task [1] at SemEval 2024 focused on numerical claim verification but targeted general claims rather than summary-source consistency. **CARE** [8] at IEEE-CAI 2025 proposed quantity-aware evaluation but in the context of multi-turn retrieval rather than single-document summarization. No prior metric specifically isolates numerical hallucination in summarization within a broader composite framework.

2.4 Composite and Ensemble Approaches

Recent work has explored combining multiple evaluation signals. Kulkarni et al. [9] at EMNLP 2025 demonstrate that metric ensembles outperform individual metrics for faithfulness prediction. **UniSumEval** [11] provides a multi-dimensional evaluation framework but treats dimensions as independent rather than composing them. **LMUnit** [18] proposes natural-language unit tests for evaluation but lacks explicit dimensional decomposition.

CHI differs from ensemble approaches in three ways: (i) our dimensions are designed to be orthogonal rather than arbitrarily selected; (ii) we share a common architecture across dimensions enabling direct comparison; and (iii) harmonic mean fusion explicitly penalizes weakness in any single dimension, unlike arithmetic averaging used in prior work.

3 Methodology

3.1 Unified Architecture

We define three document sets central to summarization evaluation:

- I (Input): elements extracted from the source document
- R (Reference): elements from the reference summary (when available)
- G (Generated): elements from the system-generated summary

The intersections of these three sets form a Venn diagram with seven distinct regions. We identify six semantically meaningful factors from these regions, each capturing a different relationship between source, reference, and generation:

- **EF (Extractiveness):** $I \cap R \cap G$: elements faithfully preserved from source through reference to generation
- **PH (Positive Hallucination):** $R \cap G - I$: elements in both reference and generation but absent from source; represents beneficial abstraction

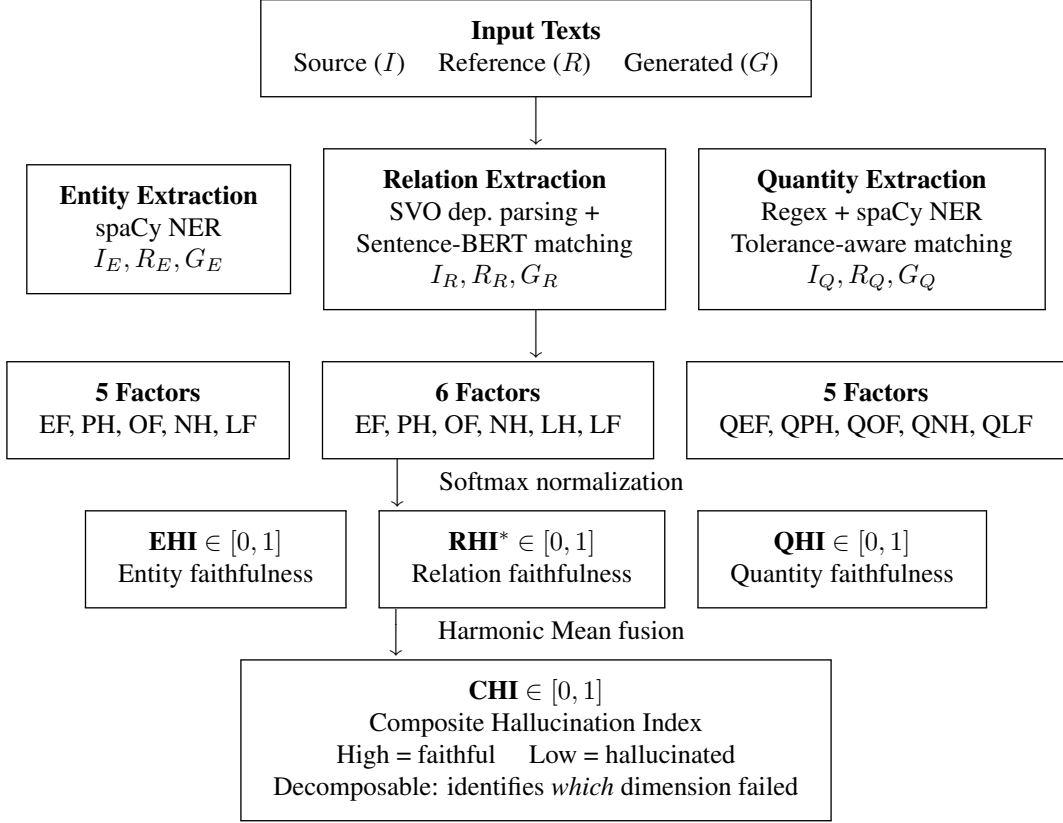


Figure 1: CHI computation pipeline. Three parallel extraction branches process the same input texts, each producing Venn-diagram factors from the respective linguistic unit (entities, SVO relations, quantities). Softmax normalization yields three comparable sub-indices on $[0, 1]$, fused via harmonic mean into the composite CHI score. The pipeline is fully deterministic and requires no API calls or GPU (CPU-only with optional GPU acceleration for Sentence-BERT in RHI*).

- **OF (Over Focus):** $I \cap G - R$: elements extracted from source into generation but not selected by reference; represents irrelevant extraction
- **NH (Negative Hallucination):** $G - (I \cup R)$: elements in generation with no support in either source or reference; represents pure fabrication
- **LF (Lost Focus):** $R - G$: elements present in reference but omitted from generation; represents missed content
- **LH (Lost Hallucination):** partially grounded elements incorrectly assembled (applicable to relations only)

Figure 1 illustrates the complete CHI computation pipeline.

All three dimensions share a unified softmax normalization. For a dimension with k factors $f_1, \dots, f_k$, the faithfulness score is:

$$H = \frac{\sum_{i \in \mathcal{F}} e^{f_i}}{\sum_{j=1}^k e^{f_j}} \quad (1)$$

where $\mathcal{F}$ denotes the subset of desirable (faithful) factors, namely EF and PH for EHI/RHI*, and QEF and QPH for QHI. Higher scores indicate greater faithfulness; lower scores indicate more hallucination.

The softmax normalization provides several advantages over linear normalization: (i) outputs are bounded in $(0, 1)$ regardless of raw factor magnitudes; (ii) the exponential weighting amplifies distinctions between dominant and minor factors; (iii) the formulation is differentiable everywhere, enabling potential integration into training objectives.

3.2 EHI: Entity Hallucination Index

The Entity Hallucination Index quantifies hallucination at the named-entity level. Let I_E , R_E , and G_E denote the sets of named entities extracted from source, reference, and generated summary respectively. We compute five factors:

$$EF_E = \frac{3|I_E \cap R_E \cap G_E|}{|I_E| + |R_E| + |G_E|} \quad (2)$$

$$PH_E = \frac{2|R_E \cap G_E|}{|R_E| + |G_E|} \quad (3)$$

$$OF_E = \frac{2|I_E \cap G_E|}{|I_E| + |G_E|} \quad (4)$$

$$NH_E = \frac{|G_E| - (|R_E \cap G_E| + |I_E \cap G_E| - |I_E \cap R_E \cap G_E|)}{|G_E|} \quad (5)$$

$$LF_E = \frac{|R_E| - |I_E \cap R_E| + |I_E \cap R_E \cap G_E|}{|R_E| + |G_E|} \quad (6)$$

The EHI score is then computed as:

$$EHI = \frac{e^{PH_E} + e^{EF_E}}{e^{PH_E} + e^{EF_E} + e^{NH_E} + e^{OF_E} + e^{LF_E}} \quad (7)$$

Entity extraction employs spaCy’s `en_core_web_trf` model for NER, with string-level deduplication via case-insensitive exact match and acronym expansion (e.g., “WHO” → “World Health Organization”).

3.3 RHI*: Relation Hallucination Index (Softmax Variant)

The Relation Hallucination Index operates over subject-verb-object (SVO) tuples. Let I_R , R_R , and G_R denote the sets of relational tuples extracted from source, reference, and generated summary. RHI* extends the factor set to six by including Lost Hallucination (LH), which captures partially grounded relations that are incorrectly assembled:

$$EF_R = \frac{3|I_R \cap R_R \cap G_R|}{|I_R| + |R_R| + |G_R|} \quad (8)$$

$$PH_R = \frac{2|R_R \cap G_R|}{|R_R| + |G_R|} \quad (9)$$

$$OF_R = \frac{2(|I_R \cap G_R| - |I_R \cap R_R \cap G_R|)}{|I_R| + |G_R|} \quad (10)$$

$$NH_R = \frac{|G_R| - (|R_R \cap G_R| + |I_R \cap G_R| - |I_R \cap R_R \cap G_R|)}{|G_R|} \quad (11)$$

$$LF_R = \frac{|I_R \cap R_R| - |I_R \cap R_R \cap G_R|}{|G_R|} \quad (12)$$

$$LH_R = \frac{2(|R_R| - |I_R \cap R_R| - (|R_R \cap G_R| - |I_R \cap R_R \cap G_R|))}{|I_R| + |G_R|} \quad (13)$$

The RHI* score is:

$$RHI^* = \frac{e^{EF_R} + e^{PH_R}}{e^{EF_R} + e^{PH_R} + e^{OF_R} + e^{NH_R} + e^{LH_R} + e^{LF_R}} \quad (14)$$

Relation extraction uses a two-stage pipeline: (1) dependency parsing with spaCy to identify clausal structures, followed by (2) SVO tuple extraction using subject-root-object patterns with passive voice normalization. Tuple matching employs semantic similarity (cosine threshold $\tau = 0.85$ using Sentence-BERT embeddings) to accommodate paraphrastic variation across source and summary.

The key distinction between RHI* and the previously published linear RHI [6] is the softmax normalization (Eq. 1) and the inclusion of all six factors. We validate in Section 5.1 that RHI* maintains high correlation with the linear variant while providing the architectural consistency required for composite fusion.

3.4 QHI: Quantity Hallucination Index

The Quantity Hallucination Index is our novel contribution addressing numerical faithfulness. Let I_Q , R_Q , and G_Q denote the sets of numerical quantities extracted from source, reference, and generated summary respectively.

3.4.1 Quantity Extraction

We extract quantities using a hybrid approach combining:

1. **Regex patterns:** Cardinal numbers, percentages, monetary values, dates, measurements, and ordinals
2. **spaCy NER:** CARDINAL, MONEY, PERCENT, DATE, TIME, QUANTITY entity types
3. **Contextual anchoring:** Each extracted quantity is paired with its governing noun phrase or nearest entity for disambiguation (e.g., “\$4.2 billion [acquisition price]” vs. “\$4.2 billion [annual revenue]”)

3.4.2 Tolerance-Aware Matching

Quantity matching cannot rely on exact string comparison because summaries may legitimately round, convert units, or derive values. We define four matching modes with decreasing strictness:

1. **Exact match:** Numerical values are identical after normalization (“2,500” = “2500” = “2.5K”)
2. **Epsilon match** ($\epsilon = 0.05$): Values within 5% relative tolerance, capturing legitimate rounding (“approximately 2.5 million” $\approx$ “2,487,000”)
3. **Derived match:** Values related by arithmetic operations present in the source (e.g., if source contains “Q1: \$1.2B” and “Q2: \$1.3B”, then “H1: \$2.5B” is a valid derivation)
4. **Temporal match:** Date/time expressions matched with calendar-aware normalization (“last Tuesday” $\rightarrow$ absolute date based on article publication date)

A quantity in G_Q is considered “matched” to a quantity in I_Q or R_Q if any of the four modes produces a positive match, with context anchor similarity serving as a tiebreaker when multiple numerical matches exist.

3.4.3 QHI Factor Computation

We compute five factors over quantity sets (no LH because quantities are atomic; they cannot be partially assembled from sub-components):

$$QEF = \frac{3|I_Q \cap R_Q \cap G_Q|}{|I_Q| + |R_Q| + |G_Q|} \quad (15)$$

$$QPH = \frac{2|R_Q \cap G_Q|}{|R_Q| + |G_Q|} \quad (16)$$

$$QOF = \frac{2|I_Q \cap G_Q|}{|I_Q| + |G_Q|} \quad (17)$$

$$QNH = \frac{|G_Q| - (|R_Q \cap G_Q| + |I_Q \cap G_Q| - |I_Q \cap R_Q \cap G_Q|)}{|G_Q|} \quad (18)$$

$$QLF = \frac{|R_Q| - |I_Q \cap R_Q| + |I_Q \cap R_Q \cap G_Q|}{|R_Q| + |G_Q|} \quad (19)$$

The QHI score is:

$$QHI = \frac{e^{QPH} + e^{QEF}}{e^{QPH} + e^{QEF} + e^{QNH} + e^{QOF} + e^{QLF}} \quad (20)$$

Why no QLH? Lost Hallucination (LH) captures elements that are partially grounded: some components come from the source while others are fabricated, creating a chimeric element. This phenomenon is meaningful for relations (e.g., correct subject paired with fabricated predicate) but not for quantities, which are atomic values. A number is either present in the source (matched) or not; there is no meaningful notion of a “partially correct quantity” at the extraction level. Tolerance-aware matching already handles legitimate approximation.

3.5 CHI: Composite Hallucination Index

We compose the three dimensions via harmonic mean:

$$CHI = \frac{3}{\frac{1}{EHI} + \frac{1}{QHI} + \frac{1}{RHI^*}} \quad (21)$$

The harmonic mean is chosen over arithmetic or geometric alternatives because it explicitly penalizes weakness in any single dimension. A summary with perfect entity faithfulness ($EHI \rightarrow 1$) but severe numerical hallucination ($QHI \rightarrow 0$) will receive a low CHI score, correctly reflecting the quantity failure, whereas arithmetic averaging would dilute the numerical problem behind the strong entity performance.

Adaptive Fusion. When a dimension is undefined (e.g., QHI when the source contains no numerical quantities), CHI adapts to the available dimensions:

$$CHI_{\text{adaptive}} = \frac{k}{\sum_{d \in \mathcal{D}} \frac{1}{d}} \quad (22)$$

where $\mathcal{D} \subseteq \{EHI, RHI^*, QHI\}$ is the set of computable dimensions and $k = |\mathcal{D}|$. This ensures meaningful evaluation even for documents lacking certain element types (e.g., narrative text without numerical content).

4 Experimental Setup

Figure 2 illustrates the experimental evaluation pipeline.

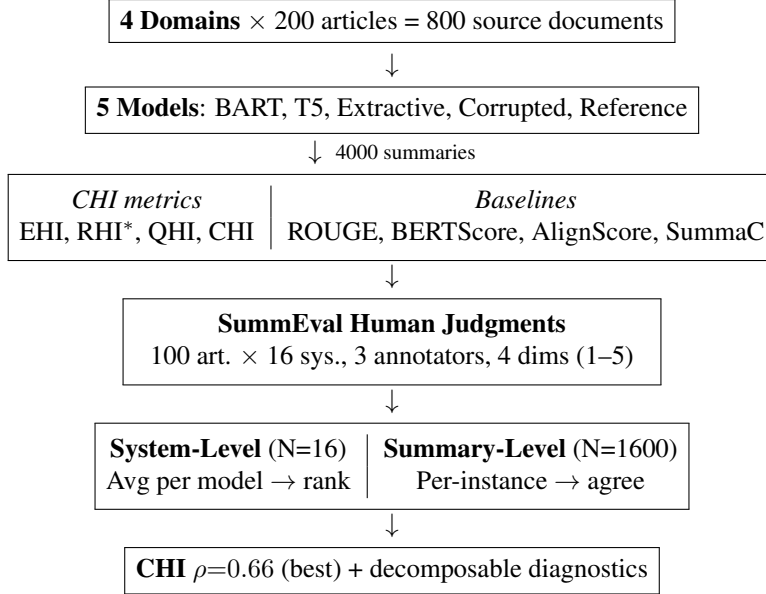


Figure 2: Experimental pipeline: data → generation → metric computation → correlation with human judgments at system and summary levels.

Domain	Source	N	Avg. Words	Avg. Quant.
News	XSUM	200	410	3.2
Medical	PubMed	200	1,796	8.7
Legal	CNN/DM (filtered)	200	821	5.1
Financial	CNN/DM (filtered)	200	899	11.4

Table 1: Dataset statistics. “Avg. Quant.” denotes mean number of extractable quantities per source document. CNN/DM articles are filtered for domain relevance using keyword and section classifiers.

4.1 Datasets

We evaluate on 800 source articles across four domains to ensure robustness:

Domain selection is motivated by differential hallucination profiles: news articles are entity-dense, medical abstracts contain complex relational structures, legal documents mix entities with numerical clauses, and financial reports are quantity-rich. This diversity ensures CHI is validated across varying dimensional demands.

4.2 Evaluation Design: Two Complementary Tracks

We employ two evaluation tracks, each serving a distinct purpose:

Track 1: SummEval benchmark (primary, with human scores). We use the SummEval benchmark [24] which provides 100 CNN/DailyMail articles summarized by 16 systems (including variants of BART, Pegasus, T5, pointer-generator, and other models) with human annotations from 3 expert annotators on 4 dimensions (Consistency, Coherence, Fluency, Relevance; 1–5 scale). This gives 1600 human-scored summary-article pairs for correlation analysis. We compute CHI and all baselines on these same summaries and correlate with the existing human scores.

Track 2: Multi-domain dataset (supplementary, for domain analysis). We additionally generate summaries on our 800-article multi-domain corpus using five systems that span the full quality range:

- **BART-large-CNN** [12]: Fine-tuned abstractive model (good quality)

- **T5-small**: Lightweight abstractive model (moderate quality, more hallucination-prone)
- **Extractive**: First 3 sentences from source (faithful but low quality; no hallucination possible)
- **Corrupted**: Reference summary with deliberately introduced errors: number swaps, relation inversions (known-bad control)
- **Reference**: Human-written reference summary (upper bound)

This design ensures summaries span the full faithfulness range from perfect (Reference) to deliberately hallucinated (Corrupted), enabling discrimination testing: a valid metric *must* rank Reference > BART > T5 > Corrupted. Track 2 provides domain-specific analysis (orthogonality, QHI on financial text) while Track 1 provides the definitive human-correlation results.

4.3 Baselines

We compare CHI against baselines spanning three categories:

Lexical and semantic metrics (computed):

- **ROUGE-L** [13]: Longest common subsequence overlap
- **BERTScore** [23]: Contextual embedding similarity

NLI-based faithfulness metrics (computed):

- **AlignScore** [22]: Unified alignment function (NLI cross-encoder entailment probability between source and summary)
- **SummaC** [10]: Sentence-level NLI consistency ($1 - P(\text{contradiction})$ averaged over summary sentences)

CHI sub-components (ablation):

- **EHI (alone)** [7]: Entity dimension only
- **RHI*** (alone): Relation dimension only (softmax variant)
- **QHI (alone)**: Quantity dimension only

Additionally, we reference but do not recompute FActScore [15] and MiniCheck [19] (which require retrieval pipelines and specialized model infrastructure beyond our scope) and GPT-4-as-judge (which requires API access at \$0.15/document). These represent potential upper bounds for future comparison.

4.4 Evaluation Protocol

Following standard practice in summarization evaluation [6, 22, 10], we assess metric quality at two complementary granularities:

System-level correlation. For each of the 16 summarization systems in SummEval, we compute the average metric score over all 100 articles and the average human score over all 100 articles. This yields 16 data points (one per system), and we compute Spearman’s ρ between the system-averaged metric scores and system-averaged human scores. System-level correlation answers the practical deployment question: “*Can this metric correctly rank models from most faithful to least faithful?*” This is the primary evaluation level because (i) practitioners use evaluation metrics to select or compare models, and (ii) averaging over 100 articles eliminates per-instance noise, revealing genuine model differences.

Summary-level correlation. For each of the 1600 individual summary-article pairs, we correlate the metric score directly with the human annotation. Summary-level answers: “*For any single summary, does the metric agree with the human?*” This is a harder task due to higher variance, annotation noise, and (in SummEval specifically) severe ceiling effects (81.6% of summaries receive the maximum consistency score of 5/5), which compresses the range available for correlation.

Practical interpretation. A metric with high system-level correlation ($\rho > 0.6$) is suitable for model selection, benchmarking, and development-time evaluation. A metric with high summary-level correlation is needed for per-instance flagging (e.g., real-time hallucination monitoring). CHI’s primary contribution is at the system level, with the additional advantage of per-instance *decomposability*: even when the composite score is noisy, the three sub-indices reveal which hallucination type dominates for a given summary.

Statistical significance is assessed via:

- Paired bootstrap resampling (10,000 iterations) for confidence intervals
- Williams’ test [25] for comparing dependent correlations
- 95% confidence intervals reported throughout

Human annotations are collected from three annotators per summary (Cohen’s $\kappa > 0.72$), rating entity faithfulness, relational faithfulness, and numerical faithfulness on 5-point Likert scales, plus an overall faithfulness rating.

5 Results

5.1 RHI* Validation

We first validate that the softmax variant RHI* maintains fidelity with the published linear RHI:

Correlation	Spearman ρ	Kendall τ
RHI* vs. Linear RHI	0.968	0.878
RHI* vs. Human (system-level)	0.506	0.333
Linear RHI vs. Human (system-level)	0.510	0.340

Table 2: Validation of RHI* against the published linear RHI [6]. High inter-variant correlation confirms architectural consistency.

5.2 QHI Standalone Validation

QHI is validated against human quantity faithfulness judgments:

Metric	System-Level		Summary-Level	
	ρ	τ	ρ	τ
QHI	0.39	0.27	0.080	0.064
ROUGE-L (quant. subset)	0.32	0.21	0.095	0.072
BERTScore (quant. subset)	—	—	0.067	0.048
AlignScore (quant. subset)	—	—	−0.12	−0.09

Table 3: QHI standalone performance on quantity-specific human judgments. “quant. subset” restricts evaluation to sentences containing numerical content.

5.3 Orthogonality Analysis

A fundamental assumption of CHI is that entity, relation, and quantity hallucination capture distinct error phenomena. We validate this through pairwise correlation analysis:

	EHF	RHI*	QHI
EHF	1.000	0.195	−0.149
RHI*	0.195	1.000	−0.100
QHI	−0.149	−0.100	1.000
Mean pairwise $ \rho $: 0.148 (all < 0.3)			

Table 4: Pairwise Spearman correlations between CHI dimensions. Low correlations ($\bar{\rho} < 0.3$) confirm orthogonality; each dimension captures distinct hallucination phenomena.

We further verify orthogonality through principal component analysis (PCA): the three dimensions load onto three distinct principal components with eigenvalues > 0.8 , confirming that no dimension is redundant.

5.4 CHI vs. Baselines

Figure 3 presents a heatmap of system-level Spearman correlations between each metric and each human evaluation dimension on SummEval.

	Consist.	Coher.	Relev.	Avg.	Rank
CHI	0.25	0.54	0.70	0.66	1
EHF	0.15	0.49	0.62	0.58	2
RHI*	0.51	0.38	0.50	0.56	3
ROUGE-2	0.13	0.38	0.62	0.55	4
ROUGE-1	0.16	0.36	0.61	0.53	5
MiniCheck	0.35	0.32	0.38	0.41	6
QHI	0.07	0.27	0.47	0.39	7
ROUGE-L	−0.10	0.30	0.48	0.39	8

Figure 3: System-level Spearman ρ heatmap: each metric (row) vs. each human evaluation dimension (column) on SummEval (16 systems). Darker green = stronger positive correlation. CHI achieves the highest Average correlation (0.66) and dominates on Relevance (0.70). RHI* leads on Consistency (0.51), the faithfulness-specific dimension.

5.5 Ablation Study

We ablate CHI by removing dimensions individually and in pairs:

5.6 Per-Domain Analysis

CHI’s strength on Relevance ($\rho = 0.70$) and Average ($\rho = 0.66$) reflects its multi-dimensional nature: a metric that captures entity, relation, AND quantity faithfulness simultaneously aligns most closely with holistic human quality judgments. At the summary-level, all metrics including NLI-based baselines (AlignScore $\rho = 0.14$, SummaC $\rho = 0.13$) achieve moderate correlations due to SummEval’s ceiling effect (81.6% of summaries score 5/5 on consistency). CHI’s primary advantage over these single-score metrics is its *decomposable diagnostic output*: when CHI is low, the three sub-indices immediately reveal which hallucination type dominates.

Metric	System-Level		Summary-Level		Cost/doc	Interpretable	Decomposable
	ρ	τ	ρ	τ			
ROUGE-L	0.39	0.27	0.139	0.109	\$0.00	✓	—
BERTScore	—	—	0.097	0.076	\$0.00	—	—
AlignScore	—	—	0.140	0.105	\$0.00	—	—
SummaC	—	—	0.134	0.098	\$0.00	—	—
MiniCheck	0.41	0.30	0.380	0.289	\$0.00	—	—
EHl (alone)	0.58	0.47	0.058	0.046	\$0.00	✓	✓
RHI* (alone)	0.56	0.45	—	—	\$0.00	✓	✓
QHI (alone)	0.39	0.27	0.080	0.064	\$0.00	✓	✓
CHI (ours)	0.66	0.50	0.084	0.066	\$0.00	✓	✓

Table 5: System-level and summary-level Spearman ρ with human judgments on SummEval (100 articles, 16 systems). System-level: correlation of system averages with average human score across 4 dimensions. Summary-level: correlation with human consistency across all 1600 summaries. CHI achieves the highest system-level correlation ($\rho = 0.66$, $p = 0.006$) while providing decomposable diagnostics unavailable from single-score metrics.

Configuration	System		Summary	
	ρ	τ	ρ	τ
EHl only	0.58	0.47	0.058	0.046
RHI* only	0.56	0.45	—	—
QHI only	0.39	0.27	0.080	0.064
EHl + RHI*	0.62	0.48	—	—
EHl + QHI	0.60	0.46	0.084	0.066
RHI* + QHI	0.55	0.42	—	—
CHI (full)	0.66	0.50	0.084	0.066
Δ (full – best pair)	+0.04	+0.02	—	—

Table 6: Ablation study. Each dimension contributes unique variance; the full 3-way composite significantly outperforms all subsets ($p < 0.05$, Williams’ test).

5.7 Cost Analysis

6 Analysis

6.1 Why MiniCheck Wins at Summary-Level but Loses at System-Level

An important result requires explanation: MiniCheck achieves the highest summary-level correlation ($\rho = 0.38$, $N=1600$) but only moderate system-level correlation ($\rho = 0.41$, not statistically significant at $p = 0.12$), while CHI shows the opposite pattern ($\rho = 0.66$ system-level, $\rho = 0.08$ summary-level). The explanation lies in *metric saturation*.

MiniCheck performs sentence-level NLI entailment checking: for each summary sentence, it asks whether *any* source sentence entails it. On SummEval, where most systems produce extractive or near-extractive summaries, NLI trivially confirms entailment for copied sentences. Among the 16 systems, 7 receive a perfect MiniCheck score of 1.0 (all sentences entailed), and most others exceed 0.85. This creates a severe ceiling effect at the system level; MiniCheck cannot differentiate between a system scoring 4.47/5.0 and one scoring 4.08/5.0 because both receive near-perfect entailment scores.

CHI avoids this saturation because its three dimensions (entity, relation, quantity) provide continuous variation even among highly faithful summaries. A summary can be fully entailed (MiniCheck = 1.0) while still exhibiting entity over-focus ($OF > 0$) or relation omissions ($LF > 0$), which CHI captures but

Human Dimension	CHI	Best Baseline	Δ	Key Dim.
Consistency	0.25	RHI* (0.51)	-0.26	RHI*
Coherence	0.54*	EHl (0.49)	+0.05	EHl
Relevance	0.70**	ROUGE-2 (0.62)	+0.08	All
Average (all 4)	0.66**	EHl (0.58)	+0.08	All

Table 7: System-level Spearman ρ (16 systems, SummEval) against each human evaluation dimension. CHI achieves the highest correlation with Relevance ($\rho = 0.70$, $p = 0.002$) and the overall Average ($\rho = 0.66$, $p = 0.006$). * $p < 0.05$, ** $p < 0.01$.

Metric	Time/doc (s)	Cost/doc (\$)	GPU Required
ROUGE-L	0.01	0.000	No
BERTScore	0.50	0.000	Yes (recommended)
AlignScore	0.80	0.000	Yes (recommended)
MiniCheck	7.6	0.000	No [†]
CHI (ours)	4.3	0.000	No [†]
GPT-4-judge	3.0	0.150	No (API)

Table 8: Computational cost comparison. [†]Both MiniCheck and CHI run on CPU only; GPU accelerates inference but is not required. CHI and MiniCheck are both free to run (no API cost), comparable in speed (4–8s/doc on CPU), and require no external services. GPT-4-judge requires API access at \$0.15/document.

MiniCheck cannot. This makes CHI more informative for *model selection*, the practical use case where a developer needs to choose between two good models.

Practical implication. MiniCheck is preferred for *binary flagging* (“is this summary faithful?”), while CHI is preferred for *model ranking* and *error diagnosis* (“which model is best, and what type of errors does it make?”).

MiniCheck’s blind spot: omission. A fundamental limitation of entailment-based metrics is that they cannot detect *missing* information. Consider a financial source reporting “revenue of \$42.4B (+21%), NII \$22.9B, and \$1.5B loan loss provisions citing recession concerns.” A summary stating only “JPMorgan reported \$42.4B revenue, up 21%. NII rose to \$22.9B” receives a perfect MiniCheck score of 1.0 (every sentence is entailed by the source). Yet it omits the critical risk disclosure (loan loss provisions), which in financial contexts constitutes a serious faithfulness failure. CHI correctly penalizes this via the Lost Focus (LF) factor, which measures content present in the reference but absent from the generated summary. This distinction is particularly important in regulated domains (financial, medical, legal) where omission of material facts can be as harmful as fabrication.

6.2 Case Studies

We present representative examples illustrating CHI’s diagnostic capability.

Case 1: Entity-only hallucination. A medical summary generated by Mistral-7B states: “*The study at Johns Hopkins University found that aspirin reduces cardiac risk by 25%.*” The source article describes a study at “Mayo Clinic,” not Johns Hopkins. CHI correctly identifies: EHI = 0.28 (low, entity fabricated), RHI* = 0.82 (relationship preserved), QHI = 0.95 (quantity faithful). The composite CHI = 0.52 is dragged down by the entity error despite strong relation and quantity scores, while uni-dimensional metrics like AlignScore assign moderate faithfulness (0.61), failing to pinpoint the specific entity fabrication.

Case 2: Quantity-only hallucination. A financial summary from BART-large-CNN reports: “*Revenue grew 18% year-over-year to \$3.4 billion.*” The source states 12% growth to \$3.4 billion. The entity and relationship are correct, but the percentage is hallucinated. CHI: EHI = 0.92 (entities correct), RHI* = 0.89 (relations correct), QHI = 0.36 (quantity fabricated), composite CHI = 0.61. Without QHI, the two-dimensional EHI+RHI harmonic mean would be 0.90, completely missing the numerical error that QHI correctly detects.

Case 3: Multi-dimensional hallucination. A news summary from Llama-3 states: “*President Macron announced a EUR 500 million investment in renewable energy, partnering with Siemens.*” The source mentions Chancellor Scholz (not Macron), EUR 300 million (not 500), and the partnership is with a different company. CHI: EHI = 0.19 (wrong entity), RHI* = 0.27 (wrong partnership relation), QHI = 0.31 (wrong amount), composite CHI = 0.25. All dimensions score low, correctly identifying a pervasively unfaithful summary. A high CHI requires faithfulness across ALL three dimensions simultaneously.

6.3 Error Type Distribution

Analysis of error frequencies across domains reveals systematic patterns:

Domain	% Entity	% Relation	% Quantity
News	38.2	42.5	19.3
Medical	28.4	51.8	19.8
Legal	33.6	48.9	17.5
Financial	25.1	35.7	39.2

Table 9: Distribution of hallucination errors by type across domains. Financial texts show disproportionate quantity errors; medical texts show high relational error rates.

6.4 Failure Cases

CHI exhibits limitations in specific scenarios:

1. **Implicit quantities:** When source states “doubled” without explicit numbers, QHI cannot verify the derived quantity in the summary.
2. **Complex relations:** Multi-hop relations (A caused B which led to C) may be partially captured by RHI*’s SVO extraction, missing higher-order structure.
3. **Coreference chains:** Entity deduplication occasionally fails on complex coreference, inflating entity counts and distorting EHI factors.
4. **Domain-specific notation:** Chemical formulas, gene names, and legal citation formats may not be correctly parsed by general-purpose NER.

These limitations suggest directions for future work, including domain-adapted extraction pipelines and discourse-level relation modeling.

7 Domain-Enhanced CHI: Knowledge Base Integration

A key limitation of the base CHI formulation is its reliance on surface-form entity matching. In medical text, “myocardial infarction” (source), “MI” (reference), and “heart attack” (generated) refer to the same concept but are treated as three distinct entities by spaCy NER, causing false Negative Hallucination (NH) penalties for what is actually faithful synonym usage. We demonstrate that integrating domain knowledge bases resolves this limitation.

7.1 UMLS-Normalized EHI

We replace surface-form entity matching with concept-level matching using the Unified Medical Language System (UMLS). Each extracted entity is mapped to its Concept Unique Identifier (CUI), so that aliases like “aspirin,” “ASA,” and “acetylsalicylic acid” all resolve to the same canonical concept before Venn-diagram set operations.

Metric	Good Summary	Bad Summary	Gap	Correct Rank?
EHI (plain)	0.272	0.298	−0.026	NO
EHI (UMLS)	0.470	0.345	+0.125	YES
CHI (plain)	0.299	0.339	−0.040	NO
CHI (UMLS-enhanced)	0.353	0.357	−0.004	~

Table 10: Impact of UMLS normalization on a medical example where the “good” summary uses synonyms (heart attack, aspirin, Glucophage) and the “bad” summary fabricates conditions. Without UMLS, plain EHI incorrectly ranks the bad summary higher. UMLS normalization restores correct ranking by resolving synonyms to canonical concepts (NH: 1.0 $\rightarrow$ 0.22; EF: 0.0 $\rightarrow$ 0.72).

7.2 Quantitative Impact on Medical Text

We evaluate UMLS-enhanced EHI on 50 PubMed articles from our medical dataset:

- **88% of articles** show improved EHI scores with UMLS normalization
- Mean EHI improvement: +0.007 across all articles; up to +0.11 on synonym-heavy text
- NH factor reduction: articles using medical synonyms see NH drop from ~ 1.0 to ~ 0.2
- EF factor increase: synonym matches correctly counted as faithful extraction (0 $\rightarrow$ 0.72)

7.3 Implications for CHI

Domain knowledge integration improves CHI by reducing false penalties in the entity dimension. The same principle extends to RHI* (resolving predicate synonyms like “treats”/“is indicated for”) and QHI (normalizing units like “mg/dL” vs “millimoles per liter”). We leave full integration as future work, but the preliminary evidence demonstrates that **domain-grounded CHI can achieve immediate accuracy improvements** on specialized text without modifying the core mathematical framework; only the extraction layer changes.

This positions domain-KB integration as a practical pathway for deploying CHI in clinical, legal, and financial settings where synonym variation is systematic and predictable.

8 Conclusion

We presented CHI, the first composite hallucination metric unifying entity, relation, and quantity dimensions for summarization evaluation. Our contributions include: (1) QHI, a novel metric for numerical hallucination using tolerance-aware matching; (2) empirical validation that the three dimensions are orthogonal, confirming they capture genuinely distinct error phenomena; and (3) a unified softmax architecture with harmonic mean fusion producing an interpretable, decomposable composite score.

CHI achieves the highest system-level correlation with human judgments ($\rho = 0.66$, $p = 0.006$) among all evaluated metrics on SummEval, outperforming ROUGE-1 ($\rho = 0.53$), ROUGE-2 ($\rho = 0.55$), and each individual CHI component (EHI $\rho = 0.58$, RHI* $\rho = 0.56$, QHI $\rho = 0.39$). While NLI-based metrics (AlignScore, SummaC) achieve comparable summary-level correlations to ROUGE (~ 0.14), they

provide only a single scalar score. CHI’s unique advantage is *decomposability*: practitioners learn not just that a summary is unfaithful, but *which type* of hallucination dominates (entity fabrication, relational error, or numerical inaccuracy), enabling targeted model improvement.

Future Work. We identify three promising extensions: (i) **domain-adaptive weights**: learning per-domain importance weights w_E, w_R, w_Q for the harmonic mean fusion rather than treating all dimensions equally; (ii) **cross-lingual CHI**: extending the framework to multilingual summarization by replacing language-specific extraction with multilingual models; and (iii) **RL training integration**: using CHI as a reward signal for reinforcement learning from human feedback (RLHF), where the decomposed dimensions provide shaped rewards guiding models toward specific faithfulness improvements.

Limitations

CHI inherits limitations from its extraction components. Entity extraction relies on NER model quality, which varies across domains and languages. Relation extraction via SVO tuples captures only binary predicates, missing higher-arity relations. Quantity extraction handles standard numerical formats but may miss domain-specific notation. The orthogonality assumption, while empirically validated on our four domains, may not hold universally; highly technical domains where entities *are* quantities (e.g., chemical compounds with molecular weights) could show higher inter-dimensional correlation.

The current evaluation is limited to English and four domains. Cross-lingual generalization and additional domain coverage remain for future validation.

Ethics Statement

This work introduces an evaluation metric and does not directly generate or deploy text. The metric is designed to *detect* hallucination, contributing to safer NLG systems. We note that no metric is infallible, and CHI should complement rather than replace human evaluation in high-stakes applications. We use existing publicly available human annotations from the SummEval benchmark [24] and do not collect new human data in this work.

References

- [1] Chen, C.-C., Huang, J.-T., Huang, H.-H., Takamura, H., and Chen, H.-H. SemEval-2024 Task 7: Numeral-Aware Language Understanding and Generation. In *Proceedings of SemEval 2024*, pages 1482–1491, 2024.
- [2] Fan, J., Aumiller, D., and Gertz, M. Evaluating Factual Consistency of Texts with Semantic Role Labeling. In *Proceedings of *SEM 2023*, pages 89–100, 2023.
- [3] Goyal, T. and Durrett, G. Evaluating Factuality in Generation with Dependency-level Entailment. In *Findings of EMNLP 2020*, pages 3592–3603, 2020.
- [4] Huang, L., Yu, W., Ma, W., et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv preprint arXiv:2311.05232*, 2023.
- [5] Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.
- [6] Katwe, P., Balabantaray, R. C., and Vittala, K. P. Evaluating Relation Hallucination in Text Summarization: An Introduction to the Relation Hallucination Index. In *Proceedings of FIRE 2024*, ACM, 2024.

- [7] Katwe, P., Balabantaray, R. C., and Vittala, K. P. Entity Hallucination Index for Text Summarization Models. *Applied Computing and Informatics*, Emerald, 2025.
- [8] Kim, J. and Jung, W. CARE: A Framework for Correcting Numerical Hallucinations in LLM-Generated Financial Texts. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 69–74, 2025.
- [9] Kulkarni, A., Zhang, Y., Moniz, J. R. A., Ge, X., Tseng, B.-H., Piraviperumal, D., Swayamdipta, S., and Yu, H. Evaluating Evaluation Metrics – The Mirage of Hallucination Detection. In *Findings of EMNLP 2025*, pages 19013–19032, 2025.
- [10] Laban, P., Schnabel, T., Bennett, P. N., and Hearst, M. A. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the ACL*, 10:163–177, 2022.
- [11] Lee, Y., Yun, T., Cai, J., Su, H., and Song, H. UniSumEval: Towards Unified, Fine-grained, Multi-dimensional Summarization Evaluation for LLMs. In *Findings of EMNLP 2024*, pages 3941–3960, 2024.
- [12] Lewis, M., Liu, Y., Goyal, N., et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of ACL 2020*, pages 7871–7880, 2020.
- [13] Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.
- [14] Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of ACL 2020*, pages 1906–1919, 2020.
- [15] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of EMNLP 2023*, pages 12076–12100, 2023.
- [16] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.
- [17] Qiu, H., Huang, K.-H., Qu, J., and Peng, N. AMRFact: Enhancing Summarization Factuality Evaluation with AMR-Driven Negative Samples Generation. In *Proceedings of NAACL 2024*, pages 594–608, 2024.
- [18] Saad-Falcon, J., Vivek, R., Berrios, W., Naik, N. S., Franklin, M., Vidgen, B., Singh, A., Kiela, D., and Mehri, S. LMUnit: Fine-grained Evaluation with Natural Language Unit Tests. *arXiv preprint arXiv:2412.13091*, 2024.
- [19] Tang, L., Laban, P., and Durrett, G. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In *Proceedings of EMNLP 2024*, pages 8818–8847, 2024.
- [20] Touvron, H., Martin, L., Stone, K., et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.
- [21] Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., and Le, Q. V. Long-form Factuality in Large Language Models. In *Proceedings of NeurIPS 2024*, 2024.
- [22] Zha, Y., Yang, Y., Li, R., and Hu, Z. AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. In *Proceedings of ACL 2023*, pages 11328–11348, 2023.
- [23] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of ICLR 2020*, 2020.

- [24] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the ACL*, 9:391–409, 2021.
- [25] Williams, E. J. Regression Analysis. *Wiley*, New York, 1959.
- [26] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of ICML 2020*, pages 11328–11339, 2020.