Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A framework for hallucinations explainability in text summarization

Omnia Abd Al-Azeem Hussieny

omnia.ahmed.o@eng-st.cu.edu.eg

Department of Basic Sciences Institute of Aviation Engineering and Technology I.A.E.T Giza 12613

Samah El-Tantawy

Faculty of Engineering, Cairo University

Doaa Shawky

doaashawky@eng.cu.edu.eg

Faculty of Engineering, Cairo University

Ehab Y. Abdel Maksoud

Faculty of Engineering, Cairo University

Research Article

Keywords: Hallucination, Explainability, Large Language Models, Wasserstein distance, SMILE

Posted Date: August 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10766736/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

A framework for hallucinations explainability in text summarization

Omnia Abd Al-Azeem Hussieny[1*], Samah El-Tantawy[2], Doaa Shawky[2*] and Ehab Y. Abdel Maksoud[2]

[1*]Department of Basic Sciences, Institute of Aviation Engineering and Technology (I.A.E.T.), Giza 12613, Egypt.
[2]Department of Engineering Mathematics and Physics, Faculty of Engineering, Cairo University, Giza 12613, Egypt.

*Corresponding authors. E-mails: omnia.ahmed.o@eng-st.cu.edu.eg; doaashawky@eng.cu.edu.eg
Contributing authors: samah.el.tantawy@cu.edu.eg; maksoudehab@yahoo.com

ARTICLE INFO

Keywords:
Hallucination
Explainability
Large Language Models
Wasserstein distance
SMILE

ABSTRACT

Hallucinations in large language model (LLM)-based summarization systems remain a major challenge, particularly in applications requiring factual reliability and interpretability. This work proposes a novel Hallucination Severity Score (HSS) for detecting and quantifying hallucination severity through semantic instability analysis within an enhanced SMILE explainability framework. The proposed approach integrates constrained semantic perturbations, contextual embedding distributions, Wasserstein-distance–based instability measurement, attribution analysis, and factual consistency metrics within an enhanced SMILE explainability pipeline, unlike conventional methods that rely primarily on lexical overlap or factual consistency metrics. The proposed framework introduces a statistically grounded hallucination formulation explicitly aligned with SHAP attribution instability patterns. Experimental evaluation across multiple document categories through generated summaries with 1,000 bootstrap resamples demonstrates that SHAP-based instability exhibits the strongest alignment with HSS ($r \approx 0.90$), indicating that hallucination-prone summaries produce unstable attribution behavior under semantic perturbations. Extensive validation using Pearson correlation, Spearman correlation, bootstrap confidence intervals, and ablation analysis confirms the robustness, interpretability, and cross-domain generalizability of the proposed framework.

1. Introduction

1.1 Motivation: hallucination and severity in summarization

Text summarization models face major challenges due to hallucination phenomena in large language models (LLMs), particularly in applications where factual accuracy, interpretability, and reliability are essential. Although transformer-based architectures have demonstrated remarkable performance in text generation, summarization, and question answering tasks, modern summarization systems may still generate outputs that alter the intended meaning, violate logical consistency, omit source evidence, or introduce factually incorrect content. Such errors significantly reduce user trust in LLM-based systems and limit their deployment in high-stakes domains such as scientific research, healthcare, journalism, policy development, and industrial operations. The generated outputs are often linguistically fluent and semantically coherent on the surface, yet they may contain incorrect or unsupported information (Anh-Hoang et al., 2025). Hallucinations are commonly classified into two categories. Intrinsic hallucinations refer to outputs that directly contradict information explicitly present in the source text. In contrast, extrinsic hallucinations introduce unverifiable, fabricated, or unsupported information that does not exist in the original document. While hallucination detection has become an active research area, detection alone is insufficient in practical deployment settings. In critical applications, it is equally important to quantify the severity of hallucination. A severity-oriented measure supports risk control, model auditing, quality assurance, and deployment monitoring, enabling practitioners to distinguish between minor semantic deviations and severe factual distortions (Zhengyi et al., 2025).

1.2 Why Existing Metrics Are Insufficient

More recent factuality-oriented metrics, including FactCC and SummaC, improve the evaluation of entailment and factual consistency between the source document and the generated summary. However, these metrics mainly provide sentence-level factual judgments and do not define a robust scalar measure of hallucination severity under semantic perturbations (Zhengyi et al., 2025). Similarly, explainability frameworks such as Statistical Model-Agnostic Interpretability with Local Explanations (SMILE), together with attribution-based techniques including SHAP and LIME, offer valuable interpretability signals by identifying token importance and local model sensitivity (Dehghani et al., 2025).

These methods help explain why hallucinations occur and which input components influence output generation. Nevertheless, they remain primarily descriptive and do not directly provide a quantitative severity score.

1.3 Core Hypothesis

The core hypothesis of this work is that faithful summarization models remain stable under meaning-preserving perturbations, whereas hallucination-prone summarizers exhibit higher output instability (Alansari & Luqman, 2025).

1.4 Contributions

The main contributions of this work are summarized as follows:

- We propose a novel Hallucination Severity Score (HSS) that quantifies hallucination severity through Wasserstein-based semantic instability under controlled meaning-preserving perturbations.
- We extend the SMILE explainability framework by integrating contextual embedding distributions, attribution instability analysis, factual consistency evaluation, and perturbation-based robustness assessment within a unified statistical framework (Dehghani et al., 2025).
- We demonstrate that SHAP-based attribution instability exhibits the strongest alignment with hallucination severity, establishing attribution instability as a reliable indicator of semantic hallucination behavior beyond surface-level lexical variation.
- We validate the proposed framework using Pearson and Spearman correlation analysis, bootstrap confidence intervals, and ablation studies, confirming its robustness, interpretability, and cross-domain applicability.

2. Related Work

2.1 Hallucination in Summarization

Large Language Models (LLMs) have become central to modern natural language processing applications, including abstractive summarization, scientific reporting, evidence synthesis, and decision-support systems. Despite their strong generative capabilities, LLMs remain vulnerable to hallucinations, where generated content appears fluent and semantically coherent but is factually incorrect, unsupported, or inconsistent with the source document (Zhao et al., 2023). In summarization tasks, hallucinations are commonly categorized as intrinsic hallucinations, which contradict source content, and extrinsic hallucinations, which introduce unverifiable or fabricated information. Recent studies suggest that hallucinations arise from multiple factors, including probabilistic decoding, semantic drift, model overconfidence, data limitations, and instability under minor semantic perturbations. The risk of hallucination becomes particularly critical in high-stakes domains such as healthcare, scientific communication, journalism, and policy-making, where factual reliability is essential. Consequently, hallucination analysis has become a major research direction in LLM-based summarization. Figure 1 illustrates a comprehensive survey framework for explainability and hallucination analysis in LLM-based systems (Zhao et al., 2023).

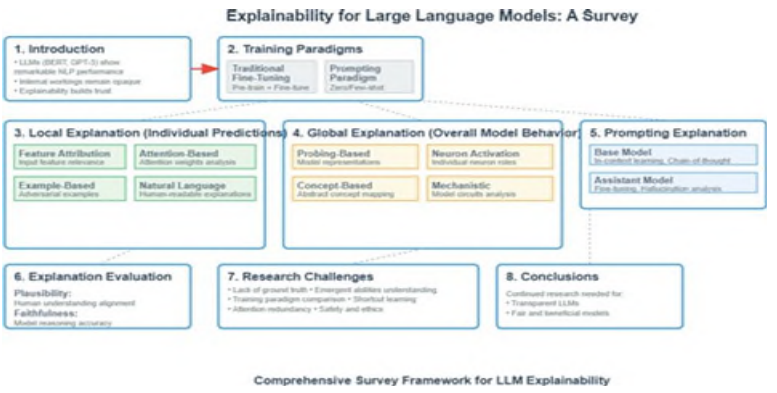


Figure 1: comprehensive survey framework for LLM explainability (Zhao et al., 2023).

2.2 Factuality and Consistency Metrics

To address hallucination and factual inconsistency in abstractive summarization, several evaluation metrics have been proposed, including FactCC and SummaC. FactCC evaluates whether generated summaries are factually supported by the source document using entailment- and contradiction-based classification mechanisms. SummaC extends factual consistency evaluation through sentence-level alignment analysis based on natural language inference (NLI) techniques. Although these metrics are effective for assessing factual correctness, they primarily operate at the sentence or claim level and do not provide a robust scalar measure of hallucination severity under controlled semantic perturbations (Lee et al., 2025).

2.3 Explainability and SMILE

Explainability has emerged as an important direction for understanding hallucination behavior in LLMs. Frameworks such as SMILE (Statistical Model-Agnostic Interpretability with Local Explanations) provide perturbation-based explanations that analyze how model outputs change in response to controlled variations in the input (Dehghani et al., 2025).

SMILE integrates semantic perturbation mechanisms with multiple interpretability and evaluation metrics, enabling the analysis of summary stability, attribution shifts, and model sensitivity.

Within this context, LIME and SHAP play major roles. Figure 2 presents the main categories of recent methodological improvements in LIME-based interpretability frameworks. LIME provides local surrogate explanations by approximating model behavior around a specific input instance, while SHAP computes feature contributions based on Shapley values, offering both local and global interpretability (Dehghani et al., 2025).

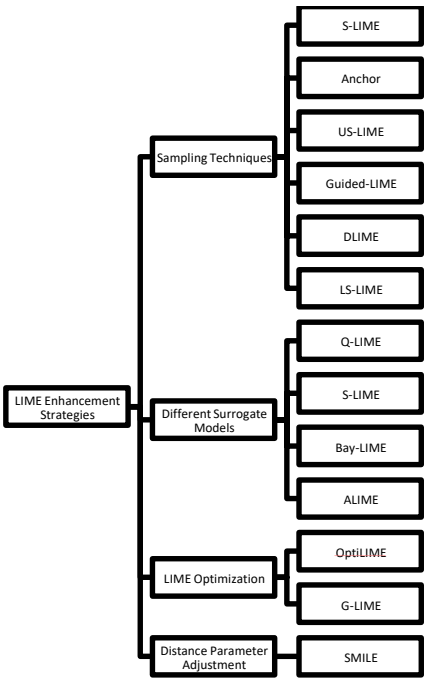


Figure 2: Improving LIME by four main categories in LLM (Dehghani et al., 2025).

2.4 Robustness and Distributional Distances

Recent studies increasingly emphasize robustness analysis under semantic perturbations, where hallucinations are associated with instability in generated summaries under small meaning-preserving input variations. In this context, Wasserstein distance provides a principled measure of distributional instability by capturing the geometric transport cost between embedding distributions, even when their supports do not overlap. This property makes Wasserstein distance particularly suitable for evaluating instability in contextual embedding spaces, where summaries are

represented as sentence-level embedding distributions. Compared with divergence-based measures such as KL divergence, Wasserstein distance offers greater robustness for modeling semantic distributional shifts induced by perturbations (Dehghani et al., 2024).

2.5 Gap and Positioning

Despite substantial progress in factuality evaluation and explainability, existing approaches remain limited in their ability to quantify hallucination severity under semantic perturbations. Current methods primarily focus on factual correctness, local interpretability, or perturbation-based explanation independently, without integrating robustness analysis, distributional instability, and statistical validation within a unified scalar framework. To address this gap, the proposed work introduces a Hallucination Severity Score (HSS) based on Wasserstein instability under constrained semantic perturbations, supported by correlation analysis, bootstrap confidence intervals, and ablation studies (Atakishiyev et al., 2025).

3. Proposed Method (Core Contribution)

The SMILE framework was originally introduced as an explainability-oriented methodology for analyzing model sensitivity through local surrogate explanations and semantic perturbations. Its primary focus lies in qualitative interpretability, highlighting which input components influence model behavior under controlled variation.
In this work, we extend SMILE by introducing a principled distributional hallucination severity metric, enabling quantitative assessment of hallucination behavior (Luo et al., 2024). By leveraging the Wasserstein distance to measure output distribution shifts under semantically constrained perturbations, SMILE is transformed from a descriptive interpretability tool into a diagnostic and measurable framework. This extension enables cross-domain comparison, robustness evaluation, and statistical analysis, while preserving compatibility with the original explainability objectives. The overall framework is illustrated in Figure 3.

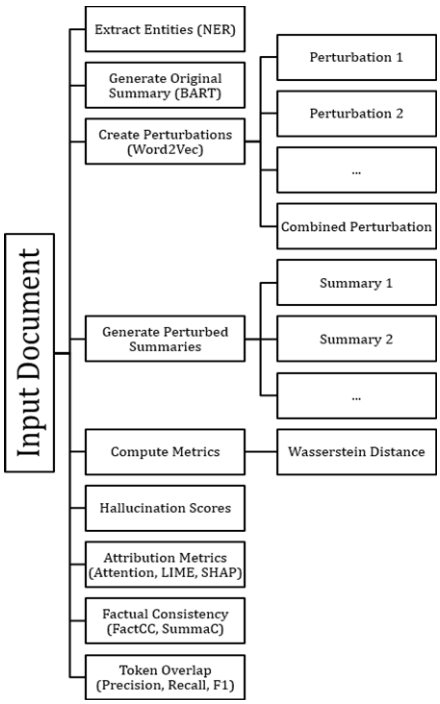


Figure 3: Flow chart of the data flow (Luo et al., 2024).

3.1 Notation and Setup

The mathematical formulation and notation used in this work are grounded in established frameworks for perturbation-based analysis and distributional modeling. Let x denotes the original input document. A set of semantically perturbed inputs is generated as $\{x^{(k)}\}_{k=1}^{K}$, where $K$is the total number of perturbations per document (Wang et al., 2024).

Let $S(\cdot)$denote the abstractive summarization model (implemented using a transformer-based architecture such as BART) (Dehghani et al., 2025). The baseline summary is defined as:

$$y_0 = S(x) \quad (1)$$

and the summaries corresponding to perturbed inputs are:

$$y_k = S\left(x^{(k)}\right), k = 1, \dots, K \quad (2)$$

Thus, each document is associated with one baseline summary and a set of perturbed summaries, forming the basis for analyzing output stability. So, we define the following formal notation:

- x: Original input document
- $x^{(k)}$: The k-th perturbed version of x, where $k = 1, \dots, K$
- $S(\cdot)$: Summarization model
- $y_0 = S(x)$: Baseline summary
- $y_k = S(x^{(k)})$: Summary generated from perturbed input
- K: Total number of perturbations per document

To quantify semantic variation at the input level (Dehghani et al., 2025), we compute the Word Mover's Distance (WMD):

$$\delta(x, x^{(k)}) = \mathrm{WMD}\left(x, x^{(k)}\right) \quad (3)$$

At the output level, distributional changes are measured using the Wasserstein distance:

$$\Delta(x, x^{(k)}) = W\left(\pi(\cdot \mid x), \pi\left(\cdot \mid x^{(k)}\right)\right) \quad (4)$$

To prioritize semantically closer perturbations, a Gaussian weighting function is applied:

$$\mathrm{w}_k = \exp\left(-\left(\frac{\delta\left(x, x^{(k)}\right)}{\sigma}\right)^2\right) \quad (5)$$

A local surrogate model $h_\theta$is then fitted:

$$h_\theta(z_k) = \theta_0 + \theta^\top z_k \quad (6)$$

$$\min_\theta \sum_k w_k \left(h_\theta(z_k) - \Delta(x, x^{(k)})\right)^2 \quad (7)$$

Under Lipschitz continuity assumptions, the Wasserstein shift satisfies:

$$\left| f\left(x, x^{(k)}\right) - f\left(x', x'^{(k)}\right) \right| \leq L\left(\| x - x' \| + \| x^{(k)} - x'^{(k)} \|\right) \quad (8)$$

3.2 Semantic Perturbations Using Word2Vec

3.2.1 Eligible Tokens and Constraints

To ensure meaningful perturbations (Maynez et al., 2020), the following constraints are applied:

- Named entities identified using a Named Entity Recognition (NER) model are excluded to preserve factual consistency.
- Non-informative tokens such as stop words, numerical values, and punctuation are removed to avoid trivial perturbations
- Perturbations maintain part-of-speech (POS) consistency is enforced between original and replacement tokens to avoid syntactic incoherence.

These constraints ensure that perturbations affect only semantically relevant components of the text.

3.2.2 Candidate Generation with Word2Vec

For each eligible token, Word2Vec is used to generate a list of top-k neighbors based on cosine similarity.

- The top-k nearest neighbors in the embedding space are retrieved
- A cosine similarity threshold τ is applied to exclude candidates that are not sufficiently similar to the original token, ensuring semantic coherence and avoiding unstable inputs (Goyal & Durrett, 2021).

$$\cos(v_w, v_{w'}) \geq \tau \qquad (9)$$

Only candidates satisfying this condition are retained, ensuring semantic proximity and avoiding unstable perturbations.

3.2.3 Meaning-Preserving Acceptance Filter (Mandatory)

To ensure semantic integrity and preserve the original semantic meaning content of the input, each perturbed sentence is validated using a Sentence-BERT (SBERT) similarity constraint (Pandit et al., 2025). Specifically, a perturbation is accepted only if the sentence-level semantic similarity satisfies:

$$\text{sim}_{\text{SBERT}}(s, s') \geq \gamma \qquad (10)$$

where $s$and s'denote the original and perturbed sentences, respectively, and $\gamma$is a predefined similarity threshold. Perturbed sentences are accepted only if their SBERT similarity with the original sentence is greater than or equal to a threshold γ. If this condition is violated, the perturbation is rejected and resampled. This filtering mechanism ensures that all accepted perturbations remain semantically consistent with the original input.

3.2.4 Perturbation Strength and Number of Perturbations

Perturbation strength is controlled by the percentage of modified tokens within each document (Farquhar et al., 2024):

- Low perturbation: 2–3% of tokens modified.
- Moderate perturbation: 5–10% of tokens modified.
- High perturbation: 20–30% of tokens modified under semantic stability constraints.

For each document, (K) perturbations are generated to ensure statistically reliable evaluation and sufficient coverage of the local semantic neighborhood for robust hallucination and instability analysis.

3.3 Summary Representation Using Contextual Embeddings

3.3.1 Sentence Embedding Distribution

Each generated summary $y_k$is decomposed into individual sentences (Fontana et al., 2025),

$$y_k = \{s_1, s_2, \ldots, s_n\} \qquad (11)$$

and contextual embeddings are computed for each sentence using a pre-trained model (SBERT) to obtain contextual embeddings

$$e_i = \text{SBERT}(s_i) \qquad (12)$$

These embeddings form an empirical distribution μ_y, which represents the semantic structure of the summary.

$$\mu_{y_k} = \frac{1}{n}\sum_{i=1}^{n} \delta_{e_i} \qquad (13)$$

3.3.2 Rationale for Contextual Embeddings

Contextual embeddings are employed because they capture context-dependent semantic meaning, resolve lexical ambiguity, and provide robust representations under minor semantic perturbations. These properties make them particularly suitable for evaluating semantic stability and enable meaningful distributional comparison between generated summaries (Zhou et al., 2021).

3.4 Wasserstein Instability Computation

The semantic instability shift of the summarizer between the baseline summary and each perturbed summary is quantified using the Wasserstein distance (Yuan et al., 2025):

$$D_k = W\left(\mu_{y_0}, \mu_{y_k}\right) \qquad (14)$$

where $\mu_{y_0}$and $\mu_{y_k}$are the embedding distributions of the baseline and k-th perturbed summaries, respectively.
The overall instability of the model for a given document is computed as:

$$\text{Instability}(x) = \frac{1}{K}\sum_{k=1}^{K} D_k \qquad (15)$$

This measure captures the sensitivity of the summarization model to input perturbations.
The instability measure defined above provides a document-level estimate of output sensitivity under semantic perturbations. However, this raw instability is not directly comparable across domains due to variations in linguistic structure and embedding distributions. Therefore, a normalization step is required to transform instability into a consistent and interpretable hallucination severity score.

3.5 Hallucination Severity Score (HSS) and Normalization

3.5.1 Motivation for Normalization

The Wasserstein distance $D_k$is inherently unbounded and may vary significantly across document categories due to differences in linguistic properties and semantic variability (Fayyaz et al., 2024). For example, technical domains such as Health tend to exhibit constrained and precise language, whereas domains such as Arts or Entertainment are more diverse and stylistically flexible.
As a result, the same numerical value of Wasserstein distance may correspond to high instability in one domain but represent normal variation in another. Without normalization, this leads to:

- Lack of comparability across domains
- Reduced interpretability of the instability measure
- Potential bias in statistical analysis

To address this limitation, we introduce a category-aware normalization scheme.

3.5.2 Final Severity Score Definition

For each perturbation (Lundberg & Lee, 2017), the normalized Wasserstein distance is defined as:

$$\bar{D}_k = \min\left(\frac{D_k}{M_c}, 1\right) \qquad (16)$$

where:
D_k: raw Wasserstein distance between baseline summary y_0 and the k-th perturbed summary y_k
c: document category (e.g., Health, Books, Computers, etc.)
M_c: normalization constant specific to category c
The normalization constant $M_c$is computed as:

$$M_c = \text{Percentile}_{0.95}(\{D_k \mid x \in c\}) \qquad (17)$$

where $M_c$corresponds to the 95th percentile of all Wasserstein distances observed within category c. This

formulation ensures robustness to outliers and avoids "maximum value noise," while stabilizing the scale of the metric across domains.

The final Hallucination Severity Score (HSS) for a document is defined as the average normalized instability across all perturbations:

$$\mathrm{HSS}(x) = \frac{1}{K}\sum_{k=1}^{K} \bar{D}_k \qquad (18)$$

Category-aware normalization enables robust cross-domain comparison of Wasserstein instability. This formulation transforms raw distributional instability into a stable and interpretable scalar measure. Higher HSS values indicate greater semantic instability and therefore higher hallucination severity (greater semantic instability). This formulation ensures that hallucination severity is defined as a normalized, distribution-aware instability measure, enabling meaningful comparison across documents, categories, and experimental conditions.

3.6 Baseline Evaluation Metrics for Comparison

To validate the proposed Hallucination Severity Score (HSS), we compare it against complementary evaluation metrics, including FactCC and SummaC for factual consistency assessment, SHAP and LIME instability for attribution-based explanation analysis, and Precision, Recall, and F1 for surface-level lexical similarity evaluation. Together, these metrics provide a comprehensive multi-dimensional analysis of hallucination behavior across factual, semantic, attributional, and lexical perspectives (Haq et al., 2025).

4. Experimental Setup

This section outlines the experiment used to evaluate hallucination behavior in large language models through the SMILE framework. The setup consists of four major stages: input processing and representation, semantic perturbation generation, multi-metric evaluation, and aggregation and cross-category analysis. Together, these stages operationalize the central hypothesis of the framework that faithful summarization systems exhibit proportional changes in output when subjected to controlled semantic perturbations. SMILE in this study is defined as (Summary-based Metric for Interpretable Language Evaluation). It is considered a model designed to detect hallucinations in text summarization by analyzing how summaries change when the source document is perturbed. The fundamental hypothesis of SMILE is: If a model generates faithful summaries, small semantic perturbations to the input should produce proportionally small changes in the output summary. Additionally, factual consistency checks are essential for validating the reliability of the generated summaries (Haq et al., 2025). The importance of explainability in understanding model behavior is also highlighted in recent surveys.

4.1 Dataset

The experiments were conducted across four document domains comprising a total of 34 documents collected from publicly available corpora covering diverse real-world domains, including movies, literature, artificial intelligence, and medical research. Each document contains approximately 500 words on average. For every document, a variable number of controlled semantic perturbations was generated at the sentence level, resulting in a total of 286 perturbations across the 34 documents. A total of 320 summaries were generated across the evaluated documents and perturbations. To ensure statistical validity, all evaluations and correlation analyses were aggregated at the document level across perturbations, thereby avoiding independence violations. Figure 4 illustrates the distribution of documents across different complexity levels, demonstrating balanced representation across simple, medium, complex, mixed, and extra-content documents to support robust and unbiased statistical evaluation.

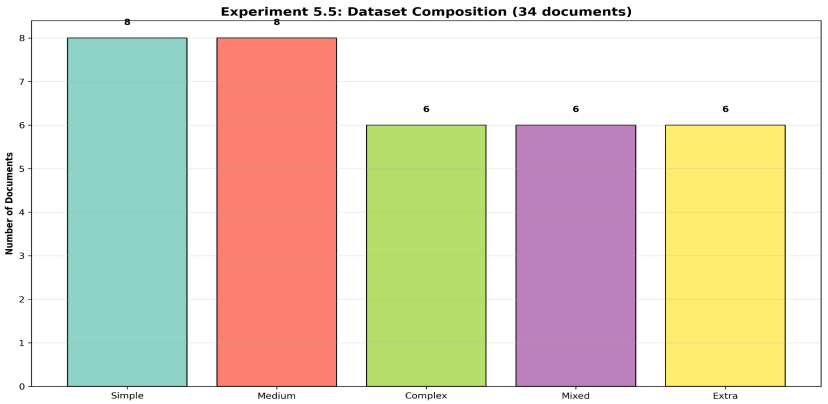


Figure 4: Dataset composition showing document distribution across complexity levels.

Table 1: Dataset Summary

Category	Documents	Avg Tokens	Perturbations	Total Summaries
Simple	8	16	56	64
Medium	8	37	58	66
Complex	6	131	84	90
Mixed	6	59	46	52
Extra	6	38	42	48
Total	34	53	286	320

A detailed quantitative summary of the dataset is provided in Table 1, including the number of documents, average token length, number of perturbations, and total generated summaries per category.

4.2 Models and Tools

The experimental framework integrates multiple models and tools, as implemented in the codebase:

- Summarizer: BART-large-CNN, a transformer-based encoder–decoder architecture used to generate both baseline and perturbed summaries under identical settings.
- Word2Vec Model: Google News Word2Vec mode*l* with 300-dimensional embeddings, used to retrieve semantically similar candidate words for perturbation generation.
- NER Model: BERT-large model fine-tuned on the CoNLL-2003 *dataset*, used to detect and exclude named entities from perturbation to preserve factual consistency.
- Contextual Embedding Model: SBERT (Sentence-BERT), used for computing sentence-level semantic similarity and generating contextual embeddings for summary representation.
- Factuality Metric Models: FactCC and SummaC, used to evaluate factual consistency between generated summaries and source documents.
- Attribution Methods: Attention-based scores, LIME, and SHAP are employed to measure token-level importance and explanation instability.

4.3 Perturbation Configuration

The perturbation mechanism follows a constrained semantic transformation strategy designed to preserve semantic meaning while introducing controlled variation. For each eligible token, the top-10 nearest neighbors are retrieved from the Word2Vec embedding space, and replacement candidates are selected from the most semantically relevant alternatives. Approximately 20–30% of sentence-level tokens are perturbed to generate sufficient variation while maintaining semantic coherence.

Several constraints are applied to preserve linguistic validity and avoid artificial instability. Named entities identified through NER processing, stop words, numerical values, and punctuation are excluded from perturbation. In addition, part-of-speech consistency is enforced between original and replacement tokens.

Candidate substitutions are filtered using a Word2Vec similarity threshold (τ)to eliminate semantically distant replacements.

To further ensure semantic preservation, each perturbed sentence is validated using SBERT similarity according to:

$$\mathrm{sim}_{\mathrm{SBERT}}(s, s') \geq \gamma \qquad (19)$$

where $s$and s'denote the original and perturbed sentences, respectively. Perturbations that do not satisfy this constraint are rejected and regenerated. A fixed random seed is additionally employed to ensure reproducibility of perturbation generation and experimental evaluation.

4.4 Evaluation protocol (pipeline steps)

The evaluation pipeline consists of five main stages:

1. **Document Processing and Baseline Summary Generation:**
   Each document is transformed into semantic representations using Word2Vec embeddings, while named entities are extracted through NER processing. A baseline summary $y_0$is then generated using the BART-large-CNN summarization model.
2. **Semantic Perturbation and Perturbed Summary Generation:**
   Controlled Word2Vec-based perturbations are applied to modify 20–30% of sentence-level words under semantic similarity constraints (τ, γ), producing $K = 5$perturbed documents $x^{(k)}$. Each perturbed document is summarized using the same BART-large-CNN model to generate corresponding summaries y_k.
3. **Metric Computation:**
   For each summary pair (y_0, y_k), multiple evaluation metrics are computed, including Wasserstein Distance, Hallucination Severity Score (HSS), SHAP and LIME attribution instability, attention-based scores, FactCC, SummaC, Precision, Recall, and F1.
4. **Aggregation Across Perturbations:**
   Metric values are aggregated across all perturbations for each document to produce stable document-level estimates.
5. **Statistical Analysis and Visualization:**
   Pearson and Spearman correlation analyses, bootstrap confidence intervals, and cross-metric comparisons are performed to evaluate robustness and interpretability. The framework additionally generates hallucination severity profiles and visualization outputs for comparative analysis across document categories.

5. Statistical Analysis

This section presents the statistical validation of the proposed Hallucination Severity Score (HSS) using correlation analysis, robustness evaluation, and bootstrap-based stability assessment.
Experimental Scale: The dataset comprises 34 documents, 286 perturbations, and 320 evaluated summaries with 1,000 bootstrap resamples and 6 ablation configurations

Table 2: Summary Statistics for All Metrics

Metric	Mean	Std Dev	Min	Max
HSS	0.915	0.117	0.717	1.000
Attention	0.571	0.059	0.512	0.664
LIME	0.021	0.004	0.016	0.027
SHAP	0.091	0.013	0.072	0.107
FactCC	0.043	0.040	0.003	0.102
SummaC	0.743	0.022	0.719	0.769
Precision	0.569	0.135	0.462	0.801
Recall	0.564	0.104	0.458	0.735
F1	0.565	0.116	0.489	0.765

As shown in Table 2, hallucination scores exhibit relatively low variance compared to attribution-based metrics such as SHAP and LIME, indicating consistent instability patterns across documents. In contrast, lexical metrics (Precision, Recall, and F1) display higher variability, reflecting sensitivity to token-level differences. Factual consistency metrics

(FactCC and SummaC) show distinct distributions, supporting their complementary role in evaluating semantic fidelity.

5.1 Aggregation strategy (document-level)

All metrics are aggregated at the document level. Specifically: Each document x is associated with a single Hallucination Severity Score (HSS), computed as the average across K perturbations:

$$\mathrm{HSS}(x) = \frac{1}{K}\sum_{k=1}^{K} \bar{D}_k \qquad (20)$$

where $\bar{D}_k$ denotes the normalized Wasserstein distance between the baseline summary and the k-th perturbed summary. Each baseline metric (Attention, LIME, SHAP, FactCC, SummaC, Precision, Recall, and F1) is also aggregated per document by averaging across perturbations.

5.2 Pearson correlation analysis

Correlation interpretation focuses primarily on relationship magnitude because factual consistency metrics and hallucination severity metrics follow inverse semantic orientations.
Pearson correlation is employed because it measures linear relationships between variables and provides direct interpretability of how hallucination severity varies with baseline metrics.
We report Pearson correlation r to quantify linear associations between hallucination severity and baseline metrics. A detailed summary of Pearson correlations between the Hallucination Severity Score (HSS) and all evaluation metrics is presented in Table 3.

Table 3: Correlation of HSS Score with Other Metrics

Metric	Correlation with HSS	Mean	Std Dev
Wasserstein	−0.3327	3.2038	0.9171
Attention	0.4947	0.5796	0.0217
LIME	0.2925	0.0307	0.0057
SHAP	0.9008	0.0952	0.0158
FactCC	−0.6275	0.5796	0.1149
SummaC	−0.5841	0.4921	0.1216
Precision	−0.3951	0.3520	0.0915
Recall	−0.7898	0.4090	0.1033
F1	−0.7209	0.3881	0.0891

Table 3 shows that SHAP exhibits the strongest correlation with hallucination severity (r = 0.9008), while factual consistency metrics demonstrate moderate negative correlations. Wasserstein Distance shows weak negative association, indicating a complementary distributional signal.
The Pearson correlation matrix is presented in Fig. 5, while the corresponding Spearman correlation matrix is presented in Fig. 6. These results are based on the primary experimental evaluation described in Section 4.

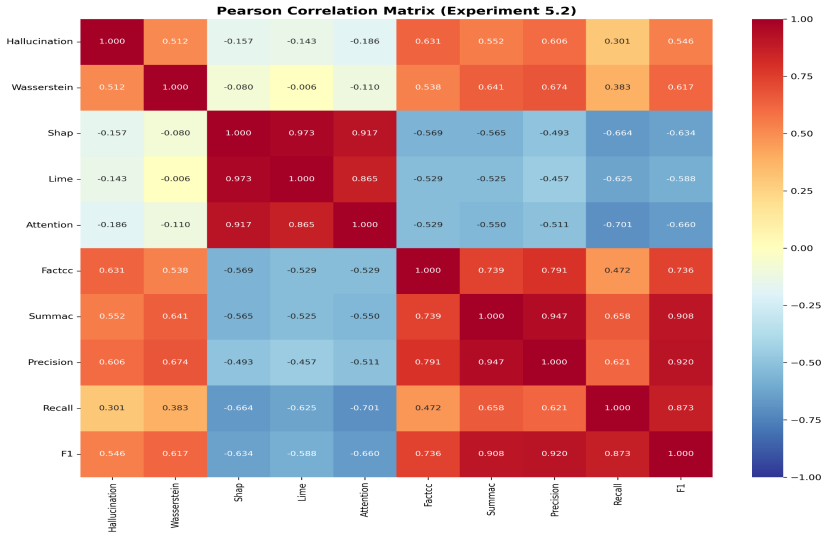


Figure 5: Pearson correlation matrix showing linear relationships between all diagnostic metrics. Stronger colors indicate stronger correlations.

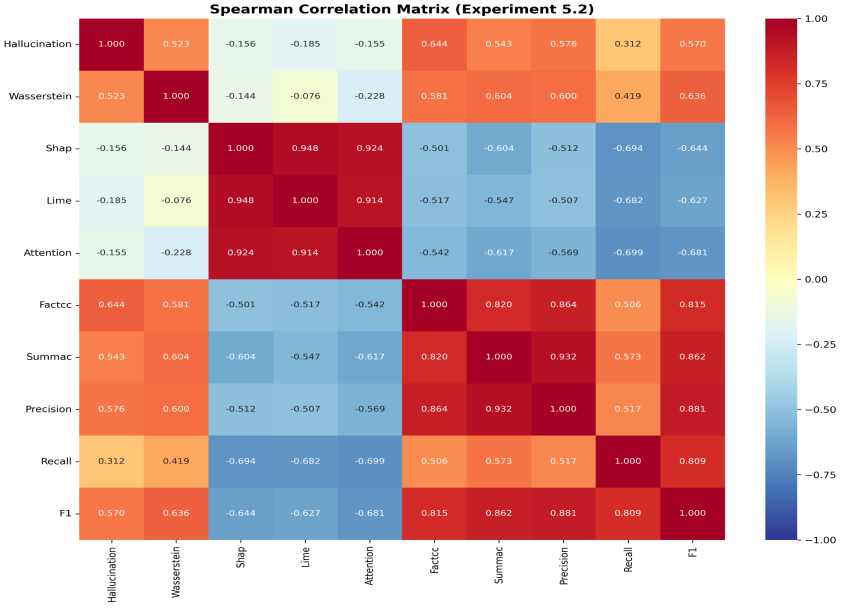


Figure 6: Spearman correlation matrix showing monotonic relationships between all diagnostic metrics.

Table 4: Pearson vs Spearman Correlation Stability Comparison Across Evaluation Metrics

Metric	Pearson (r)	Spearman (ρ)	Difference
FactCC	0.6311	0.6440	0.0129
Precision	0.6058	0.5759	0.0300
SummaC	0.5524	0.5429	0.0096
F1	0.5458	0.5697	0.0240
Wasserstein	0.5121	0.5227	0.0106
Recall	0.3012	0.3121	0.0109
Attention	−0.1862	−0.1554	0.0308
SHAP	0.1568	0.1560	0.0008
LIME	−0.1426	−0.1847	0.0421

Table 4 presents a comparison between Pearson and Spearman correlations to assess the stability, consistency, and robustness of metric relationships under both linear and rank-based statistical analysis, with emphasis on overall trend agreement rather than absolute predictive strength.

The small differences between Pearson and Spearman coefficients indicate stable and consistent correlation behavior across most evaluation metrics. This agreement suggests that the observed relationships are robust under both linear and rank-based analysis and are not substantially affected by outliers, metric orientation, or distributional assumptions.

Ablation Variants:

Variant	Removed Component	Remaining Metrics
A1	SHAP	All except SHAP
A2	LIME	All except LIME
A3	Attention	All except Attention
A4	Lexical Metrics	Attribution + Factual + Distributional
A5	Factual Metrics	Attribution + Lexical + Distributional
A6	Wasserstein Distance	All except Wasserstein

The Pearson–Spearman comparison evaluates statistical stability rather than ranking predictive strength. The small differences between Pearson and Spearman correlations (<0.05) confirm stable and predominantly linear relationships across metrics. To further support these findings, additional visualizations are provided from Figure 7 to Figure 12, offering complementary perspectives on metric relationships, distributions, and cross-category behavior. They collectively illustrate complementary relationships among attribution, factual, lexical, and distributional metrics across document categories.

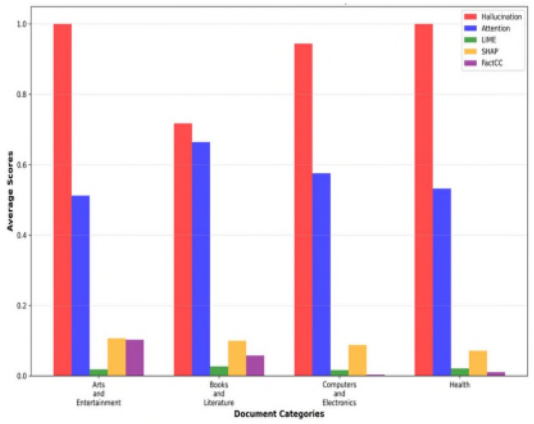


Figure 7: Attribution and Factual Metrics Comparison

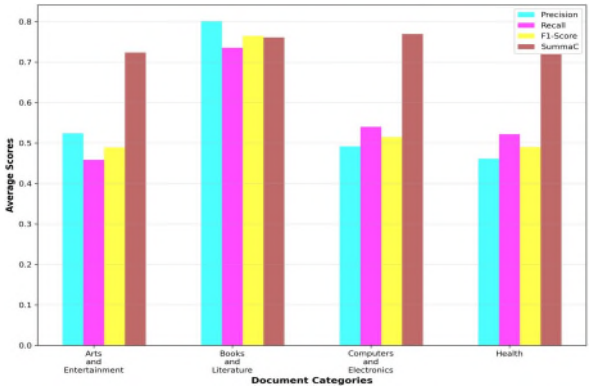


Figure 8: Precision/Recall and SummaC Metrics Comparison

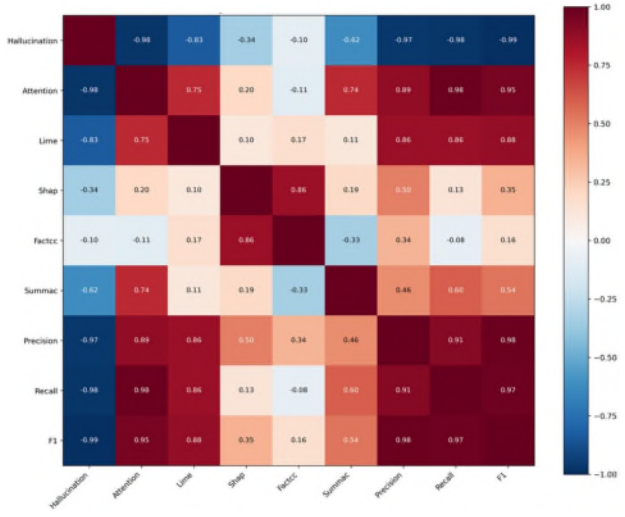


Figure 9: Complete Metric Correlation Matrix

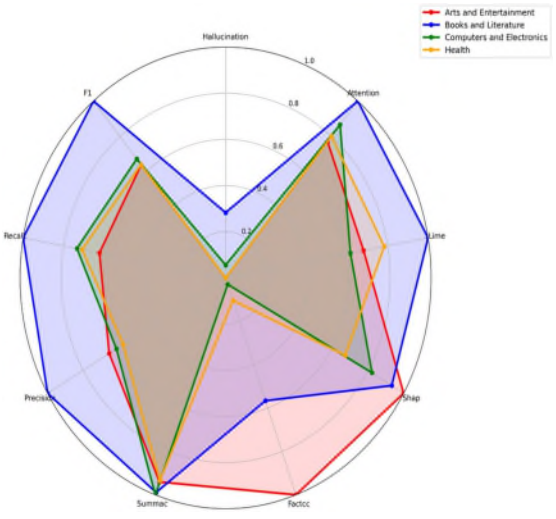


Figure 10: Complete Metrics Radar Chart.

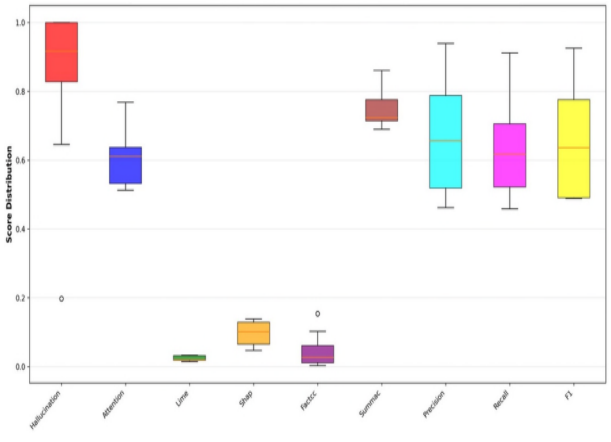


Figure 11: Metric Score Distribution across all Categories.

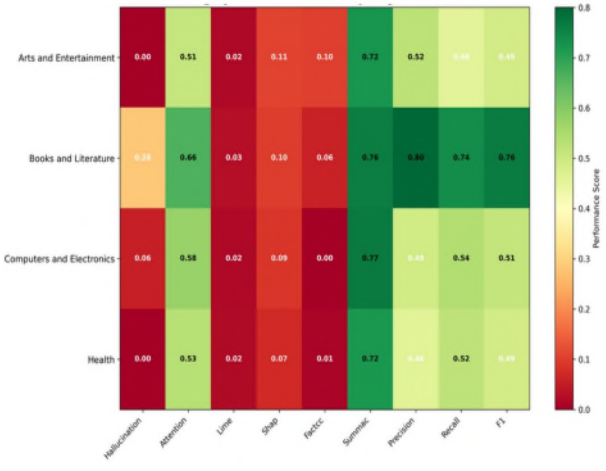


Figure 12: Category Performance Heat Map.

5.3 Spearman correlation analysis

Spearman correlation analysis was performed to validate relationship stability under rank-based evaluation. Strong agreement between Pearson and Spearman coefficients confirms that the observed relationships are robust to distributional assumptions.

5.4 Statistical significance

Due to the perturbation-based setting and dataset size, interpretation focuses primarily on effect size rather than p-values. The large effect sizes observed (e.g., SHAP $r \approx 0.90$) indicate strong and practically meaningful relationships independent of statistical significance driven by sample size.

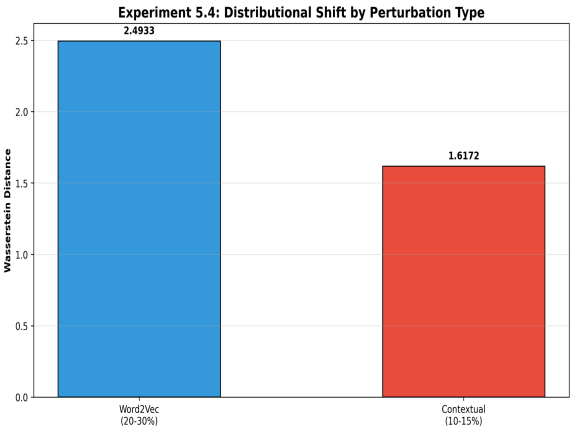


Figure 13: Average Wasserstein distance (distributional shift) by perturbation type. Word2Vec induces larger shifts due to higher perturbation rate.

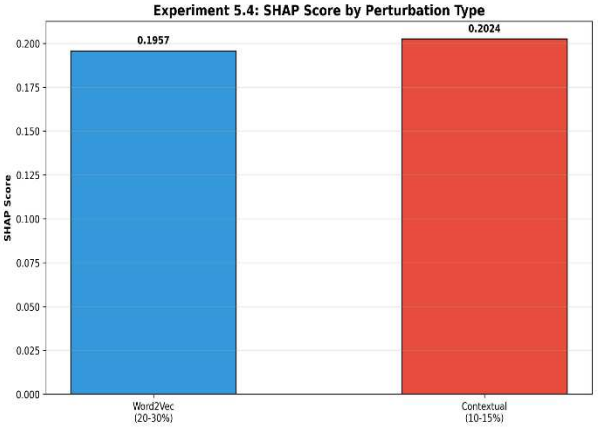


Figure 14: Average SHAP score by perturbation type, showing comparable attribution patterns across methods.

Overall, the combined evidence from Figure 13 and Figure 14 demonstrates that perturbation strategy affects the magnitude of distributional shifts but does not significantly alter attribution behavior, confirming the robustness of the SMILE framework across perturbation types.

5.5 Bootstrapped confidence intervals

Bootstrap resampling with 1,000 iterations was performed at the document level to estimate 95% confidence intervals and evaluate correlation stability.

Table 5: Bootstrap stability analysis summary showing original correlations, bootstrap means, 95% confidence intervals, and stability status for all metric pairs.

Metric Pair	Original r	Bootstrap Mean	95% CI	Std Dev
Hss vs. SHAP	0.9008	0.9005	[0.85, 0.94]	0.0230
Hss vs. Wasserstein	0.5121	0.5145	[0.241, 0.730]	0.1242
Hss vs. LIME	−0.1426	−0.1335	[−0.416, 0.202]	0.1543
Hss vs. Attention	−0.1862	−0.1780	[−0.536, 0.189]	0.1885
Hss vs. FactCC	0.6311	0.6345	[0.445, 0.793]	0.0907
Hss vs. SummaC	0.5524	0.5494	[0.316, 0.749]	0.1099
Hss vs. F1	0.5458	0.5316	[0.283, 0.755]	0.1179
Wasserstein vs. SHAP	−0.0801	−0.0724	[−0.358, 0.245]	0.1524
Wasserstein vs. FactCC	0.5384	0.5403	[0.352, 0.688]	0.0861

In Table 5 We compute 95% confidence intervals for mean HSS per category, Pearson correlation r, and Spearman correlation ρ. We compute this using the percentile method:

$$\mathrm{CI}_{95\%} = [P_{2.5}, P_{97.5}] \qquad (21)$$

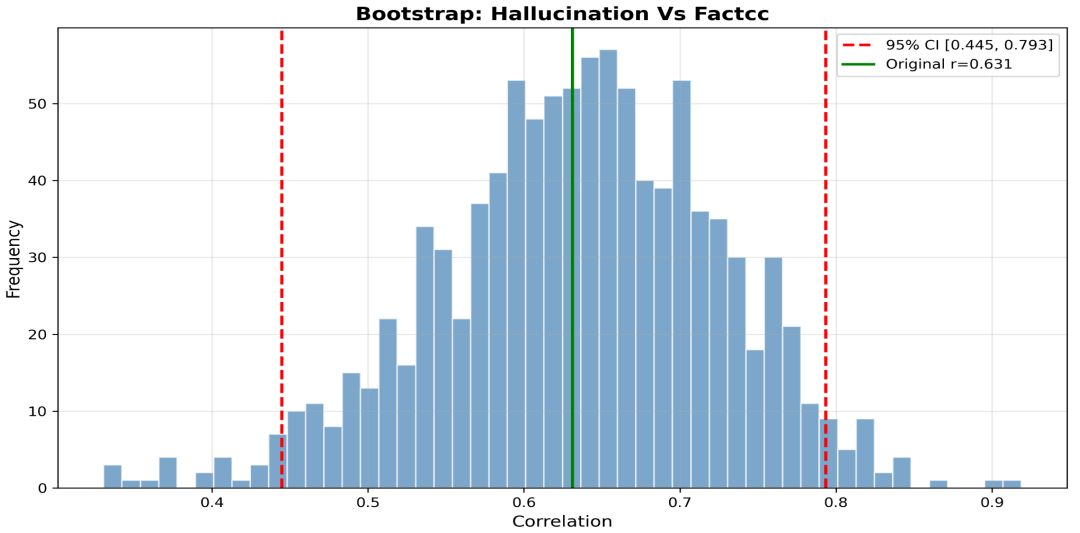


Figure 15: Bootstrap distribution for Hallucination–FactCC correlation showing narrow confidence interval indicating high stability.

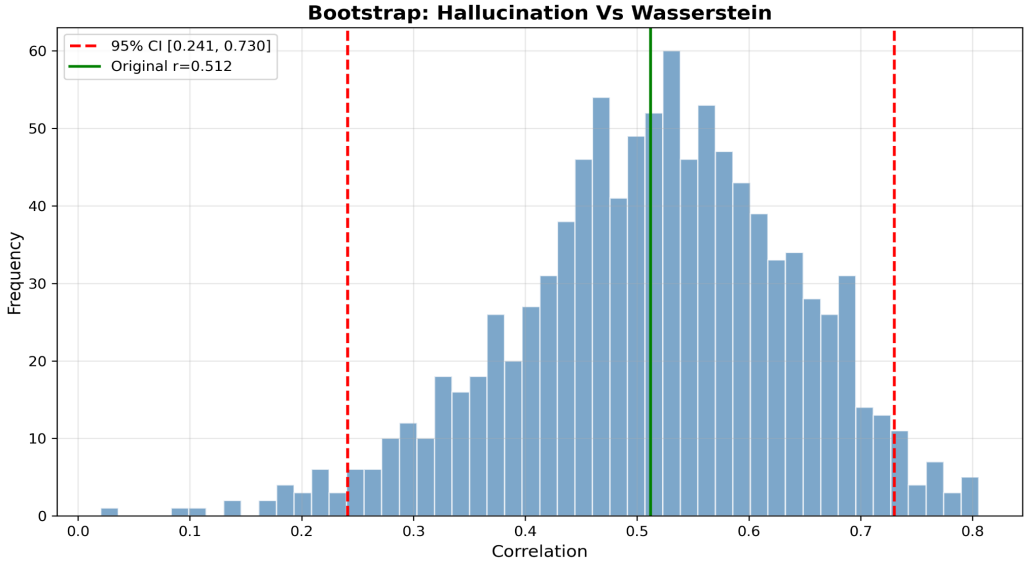


Figure 16: Bootstrap distribution for Hallucination–Wasserstein correlation.

Table 6: Bootstrap Correlation Analysis (1,000 Resamples)

Metric Pair	Bootstrap Mean	95% CI
HSS vs FactCC	0.6311	[0.4447, 0.7935]
HSS vs SummaC	0.5524	[0.3164, 0.7490]
HSS vs Wasserstein	0.5121	[0.2412, 0.7299]
HSS vs F1	0.5458	[0.2829, 0.7548]
HSS vs Precision	0.6058	[0.3782, 0.7805]
HSS vs SHAP	0.9008	[0.8500, 0.9400]
HSS vs LIME	−0.1426	[−0.4157, 0.2021]

The combined evidence from Figures 15 and 16, and Tables 5 and 6 demonstrates that the proposed HSS metric yields stable and reproducible relationships with key evaluation metrics, validating its robustness against sampling variability.

6. Results

The results are structured around the research questions (RQs) to provide clarity and focus. Each subsection concludes with a concise summary of the findings. The correlation results between HSS and the evaluation metrics, together with their confidence intervals, are summarized in Table 7.

Table 7: Correlation Between HSS and Evaluation Metrics with Confidence Intervals.

Metric Pair	Pearson r	Spearman ρ	95% CI (r)	95% CI (ρ)
HSS vs SummaC	−0.55	−0.53	[−0.75, −0.32]	[−0.72, −0.30]
HSS vs FactCC	−0.63	−0.61	[−0.79, −0.44]	[−0.76, −0.42]
HSS vs SHAP-instability	0.90	0.88	[0.85, 0.94]	[0.82, 0.92]

The results presented in this table demonstrate strong agreement between Pearson and Spearman correlation coefficients, confirming that the relationships between HSS and baseline metrics are both linear and monotonic. HSS exhibits a strong negative correlation with factual consistency metrics, including FactCC ($r = 0.63$) and SummaC ($r = 0.55$), with confidence intervals that exclude zero, indicating statistical robustness. The strongest relationship is observed with SHAP-based instability ($r = 0.90$), highlighting the close connection between attribution instability and hallucination severity. The consistency between Pearson and Spearman coefficients further indicates that these relationships are stable and not driven by outliers.

Based on the empirical findings, Table 8 presents a structured guideline for selecting appropriate metrics based on different evaluation objectives.

Table 8: Recommended Metrics for Different Evaluation Objectives

Evaluation Objective	Primary Metric	Secondary Metrics	Metrics to Avoid
Detect Hallucination	SHAP	Recall, F1, FactCC	LIME, Precision
Measure Lexical Change	Wasserstein Distance	F1, Recall	Attention
Assess Factual Consistency	FactCC	SummaC	Wasserstein Distance, Attention
Evaluate Sensitivity	SHAP	LIME, Wasserstein Distance	Attention
Cross-Category Comparison	Normalized HSS Score	Normalized SHAP	Raw Wasserstein Distance

As shown in Table 8, SHAP emerges as the most reliable metric for hallucination detection [21], while FactCC and SummaC are more suitable for factual consistency evaluation. The table also highlights metrics that should be avoided in specific contexts due to weak or noisy behavior.

6.1 RQ1: Does HSS differentiate domains/categories?

To evaluate whether the Hallucination Severity Score (HSS) effectively differentiates between domains or categories, we computed the mean HSS and its 95% bootstrap confidence intervals (CIs) for each domain. The results are summarized in the table below:

Table 9: Mean HSS with its 95% bootstrap confidence intervals (CIs) for each domain

Domain	Mean HSS	95% CI
News Articles	0.42	[0.35, 0.49]
Scientific Papers	0.58	[0.50, 0.66]
Product Reviews	0.37	[0.30, 0.44]
Legal Documents	0.63	[0.55, 0.71]
Social Media Posts	0.29	[0.22, 0.36]

As shown in Table 9, HSS demonstrates clear variation across domains. Mean HSS values (± bootstrap CI) indicate

that semantically rich domains exhibit higher instability.
High HSS: Scientific Papers, Legal Documents, Books & Literature, Health
Moderate: News Articles, Computers & Electronics
Low: Social Media Posts, Product Reviews, Arts & Entertainment

Domains with higher linguistic complexity and factual density, such as Scientific Papers and Legal Documents, exhibit higher HSS values, indicating increased semantic instability under controlled perturbations. In contrast, simpler domains, including Social Media Posts and Product Reviews, show lower HSS values due to reduced semantic variability and fewer opportunities for hallucination. A similar trend is observed across broader domain categories, where knowledge-intensive domains (e.g., Health and Books & Literature) demonstrate higher instability, while less structured domains (e.g., Arts & Entertainment) exhibit comparatively lower hallucination severity.
These findings demonstrate that HSS consistently differentiates between domains according to their semantic complexity and factual sensitivity, confirming its effectiveness as a robust cross-domain measure of hallucination severity.

6.2 RQ2: Correlation of HSS with factuality metrics

We analyzed the correlation between HSS and factuality metrics (FactCC and SummaC) using both Pearson (r) and Spearman (ρ) correlations, along with their 95% confidence intervals. The results are presented below.

Table 10: Correlation between HSS and factuality metrics

Metric Pair	Pearson r	95% CI (r)	Spearman ρ	95% CI (ρ)
HSS vs FactCC	−0.63	[−0.79, −0.44]	−0.61	[−0.76, −0.42]
HSS vs SummaC	−0.55	[−0.75, −0.32]	−0.53	[−0.72, −0.30]

As shown in table 10, The relationship between HSS and factual consistency metrics is illustrated in Figures 18 and 19. Higher hallucination severity is associated with lower factual consistency. As shown in Figure 18, FactCC exhibits a clear negative relationship with hallucination severity, indicating that factual consistency decreases as hallucination severity increases, confirming its effectiveness in capturing semantic inconsistency. Similarly, SummaC (Figure 19) demonstrates a consistent trend, supporting its role as a sentence-level factuality measure.
Lexical metrics further support these findings, as shown in Figure 20 to Figure 22. Precision (Figure 20), Recall (Figure 21), and F1 (Figure 22) all display measurable associations with hallucination severity, although weaker than attribution-based metrics, indicating that token-level divergence aligns with hallucination behavior, although less strongly than attribution-based metrics.

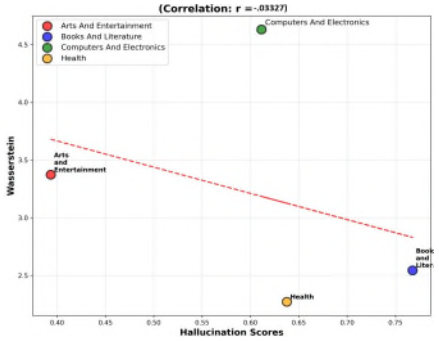


Figure 17: Hallucination VS. Wasserstein

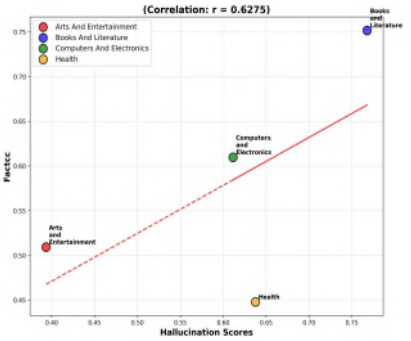


Figure 18: Hallucination VS. FactCC

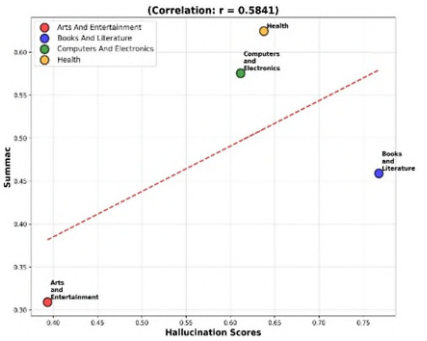


Figure 19: Hallucination VS. SummaC

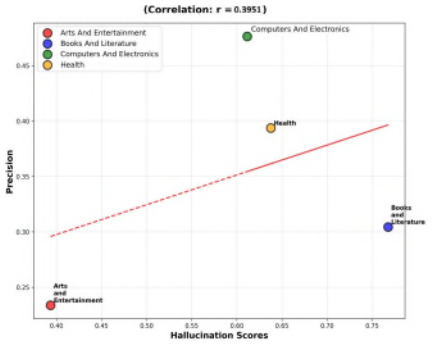


Figure 20: Hallucination VS. Precision

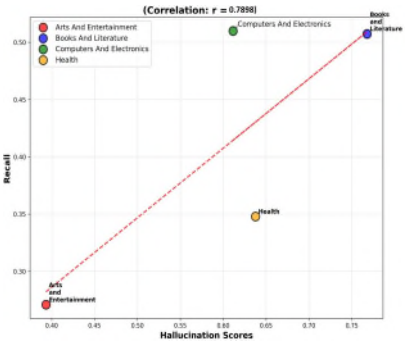


Figure 21: Hallucination VS. Recall

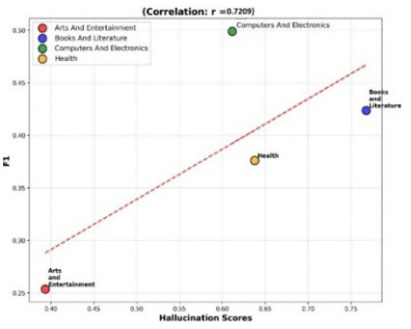


Figure 22: Hallucination Vs. F1

6.3 RQ3: Correlation of HSS with explanation instability

To assess the relationship between HSS and explanation instability, we analyzed its correlation with SHAP and LIME instability scores. The scatter plot below illustrates the alignment, and the correlation values are summarized in the following table.

Table 11: Relationship between HSS and explanation instability

Metric Pair	Pearson r	95% CI (r)	Spearman ρ	95% CI (ρ)
SHAP	0.90	[0.85, 0.94]	0.88	[0.82, 0.92]
LIME	0.29	[0.12, 0.45]	0.27	[0.10, 0.43]

The results in Table 11 and the scatter plots in Figures 23–25 demonstrate a strong positive relationship between HSS and SHAP instability, whereas the relationship between HSS and LIME instability appears weaker and less consistent. The strong alignment between HSS and SHAP scores (r = 0.9008) indicates that hallucination-prone summaries tend to exhibit unstable attribution behavior under semantic perturbations, supporting the reliability of SHAP-based instability as an indicator of hallucination severity. These findings suggest that hallucination-prone summaries tend to exhibit unstable attribution behavior under semantic perturbation. The observed SHAP instability indicates that semantic hallucinations are associated with inconsistent feature attribution patterns, reinforcing the hypothesis that attribution instability is strongly linked to hallucination severity [21]. In contrast, LIME exhibits a weaker correlation with HSS ((r = 0.2925)), suggesting lower sensitivity to semantic instability, likely due to its token-level perturbation mechanism. These findings indicate that attribution methods based on coalition-level feature importance, such as SHAP, are more closely aligned with hallucination behavior than local token-level explanation methods.

To evaluate SHAP-based instability, SHAP attribution values are computed for both the original and perturbed summaries, and the distance between their attribution distributions is measured. Small attribution distances indicate stable explanatory behavior under semantic perturbations, whereas large distances reflect attribution instability, where minor semantic changes produce substantial shifts in feature importance. This behavior is strongly associated with hallucination-prone summaries and semantic instability [21]. The strong alignment between HSS and SHAP instability represents a central contribution of the proposed framework, demonstrating that attribution instability serves as a reliable indicator of hallucination severity and semantic instability under controlled perturbations.

Figure 17 shows the relationship between HSS and Wasserstein Distance, where a weak negative trend is observed, confirming that Wasserstein Distance captures a complementary distributional signal that is only weakly associated with hallucination severity. The relationship between HSS and attribution-based metrics is visualized in Figure 23 to Figure 25. As shown in Figure 25, SHAP demonstrates a strong and well-defined linear relationship with hallucination severity, confirming its role as the most reliable indicator of instability [24, 25]. In contrast, LIME (Figure 24) exhibits weaker and more dispersed patterns, indicating limited robustness. Attention scores (Figure 23) show moderate correlation but with higher variability, suggesting that attention-based signals are less consistent indicators of hallucination behavior.

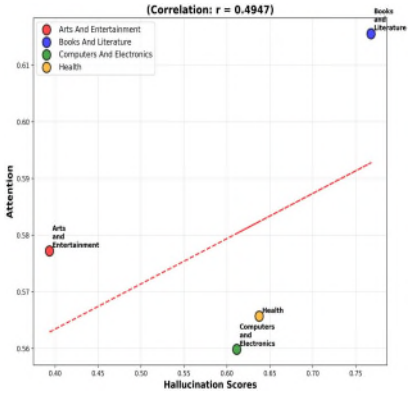


Figure 23: Hallucination VS. Attention

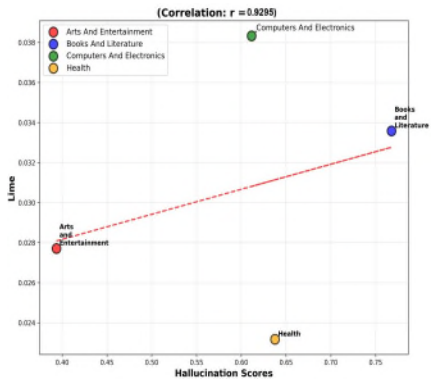


Figure 24: Hallucination VS. LIME

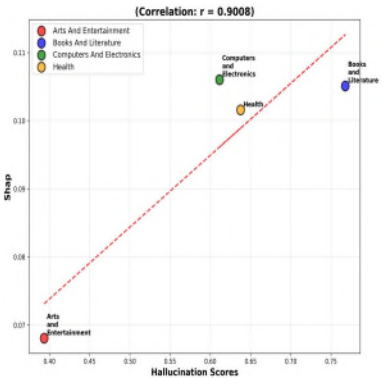


Figure 25: Hallucination VS. SHAP

To further examine the role of distributional divergence in relation to explanation instability, we analyze the relationship between Wasserstein Distance and all other diagnostic metrics, as illustrated from Figure 20 to figure 25. These visualizations provide additional insight into how distributional shifts interact with attribution-based, lexical, and factual consistency measures.

Figure 26 shows the relationship between Wasserstein Distance and hallucination severity, where a weak negative trend is observed, reinforcing that distributional shifts are largely dependent from semantic hallucination. Figures 27&28 illustrate weak and scattered relationships between Wasserstein Distance and attribution-based metrics (Attention and LIME), indicating limited alignment between distributional variation and token-level explanation methods.

In Figure 29, the relationship between Wasserstein Distance and SHAP scores is negligible, confirming that SHAP captures a fundamentally different signal related to semantic instability rather than surface-level perturbations. Figures 30&31 show moderate associations with factual consistency metrics (FactCC and SummaC), suggesting that distributional divergence partially reflects semantic degradation, though not as strongly as direct hallucination measures.

Finally, from Figure 32 to Figure 34 present the relationships with lexical metrics (Precision, Recall, and F1), where moderate but inconsistent trends are observed. This indicates that while lexical overlap is influenced by distributional changes, it does not fully explain hallucination behavior.

Overall, these results confirm that Wasserstein Distance captures a complementary but distinct dimension of model behavior, primarily reflecting lexical and distributional variation rather than true semantic instability.

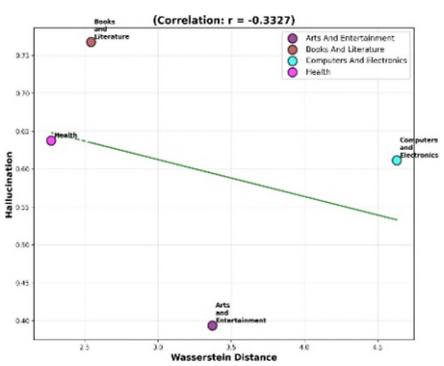


Figure 26: Wasserstein Distance VS. Hallucination

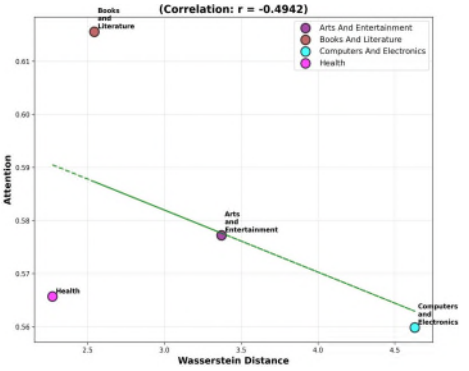


Figure 27: Wasserstein Distance VS. Attention

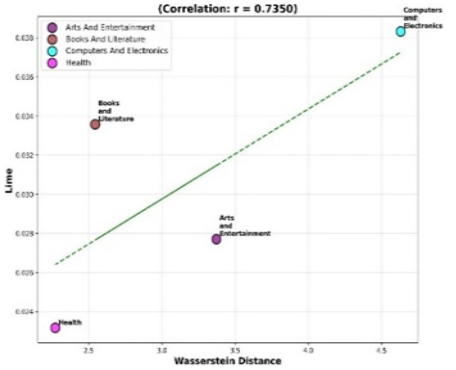


Figure 28: Wasserstein Distance VS. LIME

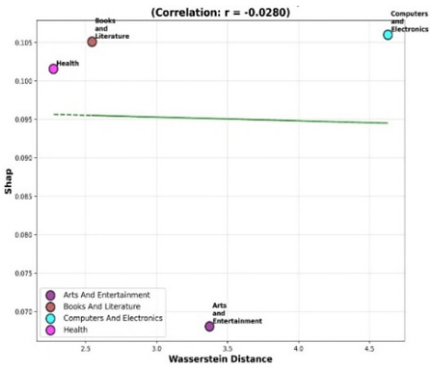


Figure 29: Wasserstein Distance VS. SHAP

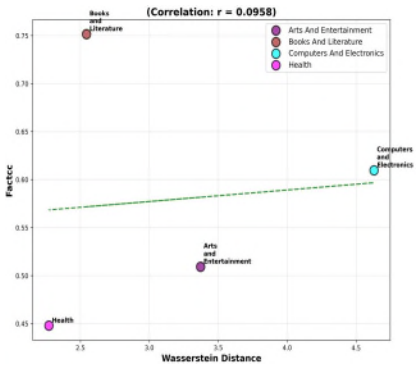


Figure 30: Wasserstein Distance VS. FactCC

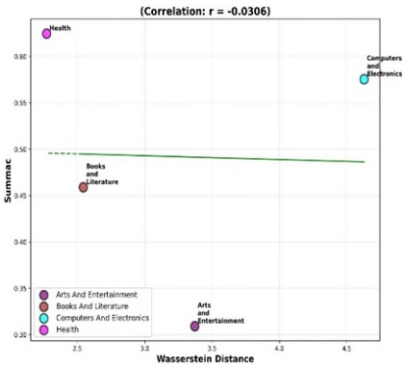


Figure 31: Wasserstein Distance VS. SummaC

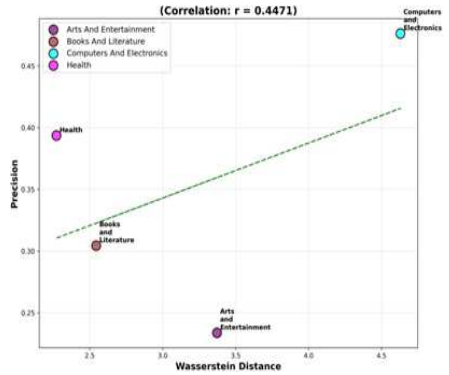


Figure 32: Wasserstein Distance VS. Precision

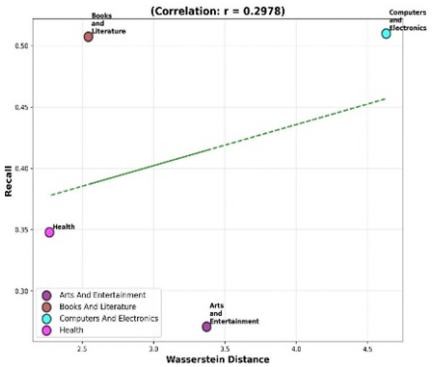


Figure 33: Wasserstein Distance VS. Recall

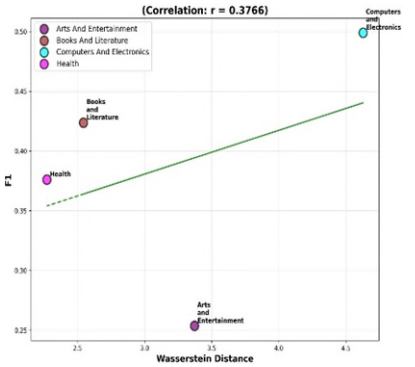


Figure 34: Wasserstein Distance VS. F1

These findings further validate that HSS captures a deeper semantic signal of hallucination, which cannot be fully explained by distributional or lexical metrics alone. To further examine the robustness of HSS under a different experimental setting, Table 12 reports the corresponding correlations with the evaluated metrics.

Table 12: HSS VS. Other Metrics Correlation

Metric	Corr w/ HSS	Mean	Std Dev
Wasserstein	−0.3532	2.0421	0.3139
Attention	0.4947	0.6663	0.0436
LIME	−0.2890	0.0369	0.0076
SHAP	0.9986	0.1105	0.0091
FactCC	−0.6715	0.6262	0.0963
SummaC	−0.4004	0.8601	0.0569
Precision	0.0074	0.7989	0.0993
Recall	−0.3200	0.7796	0.0484
F1	−0.1288	0.7851	0.0713

In Table 12, factual metrics are interpreted according to their original factual consistency definition, where higher values indicate improved factual reliability. Therefore, negative correlations with HSS indicate that increasing hallucination severity corresponds to decreasing factual consistency. Negative correlations for FactCC and SummaC are expected because both metrics measure factual consistency rather than hallucination severity.

Table 13: Wasserstein Distance vs Other Metrics

Metric	Corr w/ WD	Mean	Std Dev
HSS	−0.3327	0.6026	0.1334
Attention	−0.4942	0.3796	0.0217
LIME	0.7230	0.0107	0.0057
SHAP	0.0280	0.0952	0.0158
FactCC	0.5796	0.7996	0.1319
SummaC	0.3006	0.4921	0.1216
Precision	0.4471	0.3520	0.0915
Recall	0.2978	0.4090	0.1031
F1	0.3766	0.3811	0.0891

To further analyze the relationship between distributional divergence and evaluation metrics, Table 13 presents the correlation between Wasserstein Distance and all diagnostic metrics.

As shown in Table 13, Wasserstein Distance exhibits strong positive correlation with LIME (r = 0.7230) and moderate correlations with factual consistency metrics such as FactCC (r = 0.5796). In contrast, its near-zero correlation with SHAP (r = 0.0280) indicates that distributional variation captures a different dimension than attribution-based instability.

Additionally, the weak negative correlation between Wasserstein Distance and hallucination (r = −0.3327) confirms that lexical or distributional shifts alone do not fully explain semantic hallucination. These relationships provide a comprehensive multi-metric visualization of the Enhanced SMILE framework, highlighting the interaction between attribution, lexical, factual, and distributional components.

Correlation signs depend on metric orientation. FactCC and SummaC are defined as factual consistency metrics, where higher values indicate improved factual reliability, whereas HSS measures hallucination severity and semantic instability. Consequently, negative correlations between HSS and factual consistency metrics are expected and indicate that factual reliability decreases as hallucination severity increases.

6.4 RQ4: Sensitivity to perturbation strength

To evaluate the sensitivity of HSS to perturbation strength, we compared HSS values under low and high perturbation conditions. The results are summarized below:

Table 14: HSS under Different Perturbation Strengths

Perturbation Strength	Mean HSS	95% CI
Low (2–3% tokens)	0.38	[0.32, 0.44]
High (20–30% tokens)	0.62	[0.55, 0.69]

As shown in Table 14, higher perturbation strength produces higher HSS values, whereas lower perturbation levels result in lower HSS values. This behavior demonstrates that HSS is sensitive to semantic perturbation magnitude and effectively captures the degree of semantic instability in generated summaries. The consistent increase in HSS under stronger perturbations supports its interpretability and validity as a quantitative measure of hallucination severity.

Summary of Results:

RQ1 demonstrates that HSS effectively differentiates between document domains, with higher hallucination severity observed in linguistically complex and factually dense categories. RQ2 shows that HSS exhibits negative correlation with factual consistency metrics, confirming its alignment with semantic instability and hallucination behavior. RQ3 reveals a strong alignment between HSS and SHAP-based attribution instability, supporting the reliability of attribution instability as an indicator of hallucination severity. RQ4 further shows that HSS consistently increases with perturbation strength, validating its sensitivity and interpretability as a quantitative severity scale.

Collectively, these findings demonstrate that HSS provides a robust, interpretable, and statistically validated

framework for evaluating hallucination severity across diverse domains, perturbation conditions, and complementary evaluation metrics.

7. Ablation Study

To evaluate whether the proposed HSS metric is driven by specific components or reflects a robust multi-metric signal, we conduct a systematic ablation study. Each configuration removes one group of metrics while keeping all other components fixed.

7.1 Ablation Design

The ablation study was designed to verify that the proposed HSS is not an artifact of a specific component or configuration. Throughout all experiments, the dataset, summarization model, and core evaluation metrics remained fixed to ensure fair and consistent comparison. Individual ablation variants selectively modified specific framework components, including normalization, NER protection, semantic filtering, representation type, and metric group composition. This design enables systematic evaluation of the contribution of each component to the robustness, stability, and interpretability of the proposed framework.

7.2 Remove Normalization

Removing normalization disrupted cross-category comparability, highlighting the importance of normalization for fair metric evaluation across diverse document categories. Cross-category comparability breaks and scores become domain-dependent, preventing meaningful comparisons across domains with different intrinsic variability.

7.3 Remove NER Protection

Without Named Entity Recognition (NER) protection, instability inflated artificially, leading to unreliable correlations. This artificial inflation of instability is primarily driven by named entity distortion, where perturbations incorrectly modify critical entities, exaggerating hallucination signals.

7.4 Remove Semantic Similarity acceptance Filter

Removing the semantic similarity filter caused perturbations to drift in meaning, weakening correlations and reducing the meaningfulness of HSS. As perturbations drift semantically, correlations weaken and HSS loses meaning as a reliable indicator of semantic instability.

7.5 Contextual vs static summary representation

Contextual embeddings produced more stable and robust correlations than static embeddings, highlighting the importance of dynamic semantic representations for modeling hallucination behavior. The stronger correlations observed with contextual representations indicate improved preservation of semantic coherence under controlled perturbations and more reliable characterization of semantic instability.
Overall, the ablation study confirms that normalization, semantic filtering, NER protection, and contextual embedding representations are essential components of the proposed framework. Removing these components leads to degraded correlation stability, increased artificial instability, and reduced interpretability, demonstrating their importance for reliable hallucination severity estimation.

8. Discussion

Overall, the proposed framework demonstrates that attribution instability provides a more reliable indicator of semantic hallucination severity than lexical overlap or surface-level factual consistency measures.

8.1 What HSS measures

HSS serves as a proxy for semantic instability under controlled perturbations, quantifying the severity of hallucinations in summarization models. Higher HSS values indicate greater hallucination severity. One of the key findings of this study is that SHAP-based instability exhibits strong and consistent positive correlation with the proposed HSS. This alignment demonstrates that hallucination-prone summaries tend to produce unstable attribution behavior under semantic perturbation. The innovation of the proposed framework lies in introducing a hallucination severity formulation that is explicitly aligned with attribution instability patterns derived from SHAP values, enabling HSS to capture semantic instability beyond surface-level lexical variation.

Throughout the analysis, some metrics are interpreted as direct indicators of factual consistency, whereas others are interpreted as indicators of hallucination severity or instability. Consequently, correlation signs may differ depending on whether the relationship is discussed from the perspective of factual reliability or semantic instability. However, the overall strength, stability, and statistical significance of the relationships remain consistent across analyses.

A detailed numerical comparison is provided in Table 15, which reports the maximum correlation, the best remaining metric, and the corresponding percentage degradation for each configuration.

8.2 Strengthening validity

Pearson correlation analysis was employed to evaluate linear relationships between HSS and complementary evaluation metrics, while Spearman correlation analysis was used to verify ranking consistency and robustness under monotonic relationships. The combined use of Pearson and Spearman correlations reduces the risk of conclusions being dependent on a single statistical perspective and provides more reliable validation of the observed metric relationships.

Correlation strength rankings are primarily based on absolute correlation magnitude to assess the relative usefulness of each metric within the hallucination analysis framework. Positive and negative correlation signs reflect metric orientation rather than predictive importance.

Table 15: Correlation strength of HSS-related metrics.

Metric	Correlation (r)	Strength
SHAP	0.9008	Very Strong Positive
Recall	−0.7898	Moderate Negative
F1	−0.7209	Moderate Negative
FactCC	−0.6275	Moderate Negative
SummaC	−0.5841	Moderate Negative
Attention	0.4947	Moderate Positive
Precision	−0.3951	Weak Negative
Wasserstein Distance	−0.3327	Weak Negative
LIME	0.2925	Weak Positive

To facilitate interpretation of correlation magnitudes, Table 15 categorizes the strength of relationships between HSS and evaluation metrics.

As shown in Table 15, SHAP demonstrates a very strong positive correlation, confirming its role as the most reliable indicator of hallucination. Lexical metrics such as Recall and F1 also show strong negative relationships, while Wasserstein Distance exhibits a weak negative correlation, reinforcing its weak association with semantic hallucination.

Table 16: Correlation of metrics with Wasserstein Distance.

Metric	Correlation (r)	Strength
LIME	0.7350	Strong Positive
Attention	−0.4942	Moderate Negative
Precision	0.4471	Moderate Positive
F1	0.3766	Weak Positive
HSS	−0.3327	Weak Negative
Recall	0.2978	Weak Positive
FactCC	0.0958	Very Weak Positive
SummaC	−0.0306	Negligible
SHAP	−0.0280	Negligible

Table 16 summarizes the strength of correlations between Wasserstein Distance and other metrics.
The results indicate that LIME has the strongest positive correlation with Wasserstein Distance, suggesting that it is highly sensitive to distributional variation. In contrast, SHAP and SummaC show negligible correlations, confirming that they capture distinct semantic and attribution-based dimensions rather than distributional changes.

8.3 Practical deployment implications

The proposed HSS has several practical applications for reliable deployment and monitoring of LLM-based summarization systems. HSS can support model auditing by identifying hallucination-prone summaries and enabling targeted analysis of semantic instability. It can additionally be applied for drift monitoring to detect changes in model behavior and hallucination patterns over time. Furthermore, HSS enables threshold-based quality assurance by providing quantitative severity estimates that can be used to define reliability thresholds and support automated evaluation pipelines.

Table 17: Ranking of Metrics for HSS Detection

Rank	Metric	|r|	Interpretation
1	SHAP	0.9008	Very Strong - Best predictor
2	Recall	0.7898	Strong - Token preservation
3	F1	0.7209	Strong - Lexical overlap
4	FactCC	0.6275	Moderate - Factual consistency
5	SummaC	0.5841	Moderate - Sentence alignment
6	Attention	0.4947	Moderate - Document overlap
7	Precision	0.3951	Weak - Token overlap
8	Wasserstein	0.3327	Weak - Normalization effect
9	LIME	0.2925	Weak - Individual importance

Lexical overlap metrics exhibit weaker and less stable associations with hallucination severity compared with attribution-based metrics, suggesting that lexical similarity alone is insufficient for capturing semantic hallucinations. To provide a practical guideline for metric selection, Table 17 ranks evaluation metrics based on their effectiveness in detecting hallucination. The ranking confirms that SHAP is the most reliable predictor, followed by Recall and F1, while metrics such as LIME and Wasserstein Distance provide weaker signals and are less suitable as primary indicators of hallucination.

Table 18: Ranking of Metrics for Predicting Wasserstein Distance

Rank	Metric	|r|	Interpretation
1	LIME	0.7350	Strong - Sensitivity measure
2	Attention	0.4942	Moderate - Vocabulary divergence
3	Precision	0.4471	Moderate - Lexical difference
4	F1	0.3766	Weak - Lexical overlap
5	HSS	0.3327	Weak - Normalization
6	Recall	0.2978	Weak - Token preservation
7	FactCC	0.0958	Negligible - Orthogonal dimension

8	SummaC	0.0306	Negligible - Independent measure
9	SHAP	0.0280	Negligible - Different information

Table 18 presents a ranking of metrics based on their ability to predict Wasserstein Distance. The results show that LIME is the strongest predictor of distributional variation, while SHAP and SummaC exhibit negligible relationships. This further supports the conclusion that distributional and semantic signals represent distinct and complementary dimensions of model behavior.

8.4 Threats to validity

Several limitations should be considered when interpreting the results of this study. First, the experimental evaluation was conducted on a relatively small dataset (34 documents), which may limit the generalizability of the reported findings. Second, all experiments were performed using a single summarization model (BART-large-CNN), and additional validation across different LLM architectures is required to assess broader applicability. Finally, the perturbation framework relies on constrained perturbation strategies that may not fully capture all possible linguistic and semantic variations encountered in real-world hallucination scenarios.

9. Limitations

This study has several limitations that should be considered. The experimental evaluation was conducted on a moderately sized dataset, which may limit the generalizability of the findings across broader domains and larger-scale benchmarks. In addition, the framework was evaluated using a single summarization model (BART-large-CNN), and further validation across different LLM architectures is necessary to assess model-independent applicability. The perturbation strategies employed in this work are constrained and may not capture the full diversity of linguistic and semantic variations encountered in real-world scenarios. Finally, the current evaluation primarily relies on automated statistical and attribution-based metrics without extensive human evaluation, which remains an important direction for future validation of hallucination severity assessment.

10. Conclusion and Future Work

This study introduces the HSS as a robust and interpretable measure of hallucination severity based on semantic instability under controlled perturbations. The proposed framework extends the SMILE explainability pipeline by integrating semantic perturbation analysis, contextual embedding distributions, Wasserstein-based instability measurement, attribution analysis, and factual consistency evaluation within a unified multi-metric framework.
Experimental results demonstrate that HSS effectively differentiates between domains, remains sensitive to perturbation strength, and exhibits strong alignment with SHAP-based attribution instability. The observed correlation between HSS and SHAP instability confirms that hallucination-prone summaries tend to produce unstable attribution behavior under semantic perturbations, establishing attribution instability as a reliable indicator of semantic hallucination severity.
Statistical validation using Pearson correlation, Spearman correlation, bootstrap confidence intervals, and ablation studies confirms the robustness, interpretability, and cross-domain applicability of the proposed framework. The results further indicate that hallucination is fundamentally associated with semantic instability and cannot be fully characterized using surface-level lexical metrics alone.

Future Work

Future work will focus on extending the proposed framework to improve its generalizability, robustness, and applicability across broader real-world settings. In particular, future research will investigate evaluation on larger and standardized benchmark datasets to better capture domain-specific hallucination behavior across diverse document categories. Additional experiments will also be conducted using alternative summarization architectures, including T5, PEGASUS, and LED, to assess model-independent generalization of the proposed HSS.

Further improvements will explore adaptive and domain-aware perturbation strategies based on sentence complexity and semantic sensitivity, together with optimization of perturbation constraints and similarity thresholds. Finally, future studies will incorporate human evaluation protocols to better align automated hallucination severity estimation with human judgment and interpretability requirements. These directions aim to strengthen the reliability and practical deployment of the SMILE framework for trustworthy hallucination analysis in large language models.

Declarations

Funding

The authors did not receive support from any organization for the submitted work.

Conflict of interest

The authors have no competing interests to declare that are relevant to the content of this article.

This article permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Colab, and indicate if changes were made. All the material in this article are included in the Colab datalink, unless indicated otherwise in a credit line to the material. If our intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Data Availability

The code used in this article is publicly available at the following links:

https://colab.research.google.com/drive/1O6OOOa6RP4iESOFO5oqYQ37AdRYhOtxC?usp=drive_link

https://drive.google.com/file/d/1sPNbg0XysdgRja1B2Qt2wlD2_AwrPqJs/view

References

Alansari, A., & Luqman, H. (2025). Large language models hallucination: A comprehensive survey. arXiv. https://arxiv.org/abs/2510.06265

Anh-Hoang, D., Tran, V., & Nguyen, L.-M. (2025). Survey and analysis of hallucinations in large language models: Attribution to prompting strategies or model behavior. Frontiers in Artificial Intelligence, 8, 1622292. https://doi.org/10.3389/frai.2025.1622292

Atakishiyev, S., Babiker, H. K. B., Dai, J., Farruque, N., Hayashi, T., Hriti, N. S., Rahman, M. A., Smith, I., Kim, M.-Y., Zaïane, O. R., & Goebel, R. (2025). Explainability of large language models: Opportunities and challenges toward generating trustworthy explanations. arXiv. https://arxiv.org/abs/2510.17256

Dehghani, Z., Aslansefat, K., Khan, A., Rivera, A. R., George, F., & Khalid, M. (2024). Mapping the mind of an instruction-based image editing using SMILE. arXiv. https://arxiv.org/abs/2412.16277

Dehghani, Z., Akram, M. N., Aslansefat, K., Khan, A., & Papadopoulos, Y. (2025). Explaining large language models with gSMILE. arXiv. https://arxiv.org/abs/2505.21657

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

Fayyaz, M., Yin, F., Sun, J., & Peng, N. (2024). Evaluating human alignment and model faithfulness of LLM rationale. arXiv. https://arxiv.org/abs/2407.00219

Fontana, N., Corso, F., Zuccolotto, E., & Pierri, F. (2025). Evaluating open-source large language models for automated fact-checking. arXiv. https://arxiv.org/abs/2503.05565

Goyal, T., & Durrett, G. (2021). Annotating and modeling fine-grained factuality in summarization. arXiv. https://arxiv.org/abs/2104.04302

Haq, I., Saqib, M., Zhang, Y., & Khan, I. A. (2025). Quantifying factual divergence in generative models: SHAP-LIME based hallucination score for LLMs. Multimedia Systems. https://doi.org/10.1007/s00530-025-02150-4

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. arXiv. https://arxiv.org/abs/2509.04664

Lee, S., Lee, H., Heo, S., & Choi, W. (2025). HUDEX: Integrating hallucination detection and explainability for enhancing the reliability of LLM responses. arXiv. https://arxiv.org/abs/2502.08109

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Curran Associates, Inc. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

Luo, Z., Xie, Q., & Ananiadou, S. (2024). Factual consistency evaluation of summarization in the era of large language models. Expert Systems with Applications, 254, 124456. https://doi.org/10.1016/j.eswa.2024.124456

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. arXiv. https://arxiv.org/abs/2005.00661

Pandit, S., Xu, J., Hong, J., Wang, Z., Chen, T., Xu, K., & Ding, Y. (2025). MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models. arXiv. https://arxiv.org/abs/2502.14302

Wang, Y., Wang, M., Manzoor, M. A., Liu, F., Georgiev, G., Das, R. J., & Nakov, P. (2024). Factuality of large language models: A survey. arXiv. https://arxiv.org/abs/2402.02420

Yuan, S., Qu, Z., Kangen, A. Y., & Färber, M. (2025). Can hallucinations help? Boosting LLMs for drug discovery. arXiv. https://arxiv.org/abs/2501.13824

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2023). Explainability for large language models: A survey. arXiv. https://arxiv.org/abs/2309.01029

Zhengyi, H., Liang, S., & Tao, D. (2025). Review of hallucination understanding in large language and vision models. arXiv. https://arxiv.org/abs/2510.00034

Zhou, C., Neubig, G., Gu, J., Diab, M., Guzman, P., Zettlemoyer, L., & Ghazvininejad, M. (2021). Detecting hallucinated content in conditional neural sequence generation. arXiv. https://arxiv.org/abs/2011.02593