Interpretable Continuous Exaggeration Scoring for News Summaries

Keisuke Iwamoto[1] Kazutaka Shimada[2]

Abstract: We propose an interpretable method to assign continuous exaggeration scores to news summaries. First, we construct an exaggeration ranking of article–summary pairs using LLM-based relative judgments with a comparison-based sorting procedure. Next, we convert the ranking into scores by recursively identifying semantic midpoints to approximate an equal-interval scale and interpolating between anchor points with PCHIP, while preserving monotonicity. Robustness is improved via ensembling across multiple runs. Experiments on 1,000 Newsroom pairs show generally stable scoring, and a validity test using artificially exaggerated summaries yields higher scores than non-exaggerated ones.

Keywords: Exaggeration Scoring, Continuous Scoring, Large Language Models, Relative Evaluation

1. Introduction

In recent years, with the widespread use of social media, an environment has emerged in which opinions and interpretations of news articles are shared almost instantly. In particular, on social media platforms that impose strict character limits, short summaries or headlines are often used instead of the full article text, and these short texts tend to be shared independently. Such summaries may be written by the original authors of the articles. However, in many cases they are created by third parties or automatically generated systems. This raises a concern that some of these summaries include exaggerated expressions that overly emphasize certain parts of the original article. Even when a summary does not directly contradict the content of the article, exaggeration can give readers an excessively strong impression of the overall content. Furthermore, when opinions or interpretations based on such exaggerated summaries are shared secondarily on social media, misunderstandings of the original article may spread more widely.

The problems caused by exaggerated summaries are not limited to simple factual inaccuracies. By emphasizing specific information, exaggerated expressions often appeal strongly to readers' emotions. Previous studies have shown that, on social media, emotionally charged or sensational content is more likely to spread than information presented in a neutral manner [1], [2]. As a result, exaggerated summaries may circulate more widely than accurate ones, promoting the spread of information that highlights particular interpretations. For this reason, the diffusion of exaggerated summaries can be considered one of the factors that distort readers' understanding of news content.

[1] Department of Creative Informatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820–8502, Japan
[2] Department of Artificial Intelligence, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820–8502, Japan

Against this background, this study aims to suppress misunderstandings caused by exaggerated summaries and to prevent the formation of public opinion based on such misunderstandings. To achieve this goal, we seek to develop a method that can capture how exaggerated a summary is. In the fast-paced environment of social media, it is important to grasp the degree of exaggeration quickly and intuitively. Therefore, representing exaggeration as a quantitative indicator is considered to be an effective approach. While qualitative explanations can convey detailed information about how exaggeration appears in a summary, they require time to read and impose a high cognitive load on users, making them less suitable for social media environments. In contrast, numerical indicators are easy to perceive visually and allow readers to immediately judge whether caution is needed.

Because a large amount of information circulates on social media, it is unrealistic to rely on manual inspection for each summary. Thus, it is desirable to automatically compute such indicators. In this study, we construct a dataset in which numerical exaggeration scores are assigned to pairs of news articles and their summaries. This dataset is intended to be used in the future for training machine learning models that automatically estimate the degree of exaggeration. Figure 1 shows the overall process of constructing the exaggeration score dataset and its intended usage. This study proposes a method for building a dataset with exaggeration scores for news summaries and investigates how large language models (LLMs) can be used to assign these scores in a stable and consistent manner.

2. Related Work

2.1 Factual Consistency Evaluation in Summarization

In research on news article summarization, many studies have focused on evaluating whether a summary is consistent with the content of the original article. FactCC [3] treats factual con-

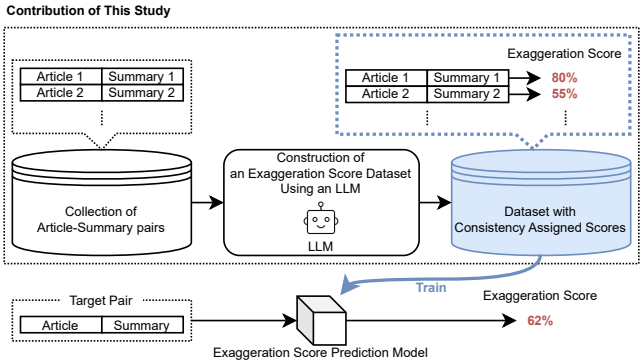

Fig. 1 Overview of this study.

sistency as a classification problem. It detects factual errors by judging whether the meanings of the source article and the summary are aligned. BERTScore [4] measures semantic similarity between the source text and the summary using contextualized embeddings from a pre-trained language model. BARTScore [5], on the other hand, evaluates the plausibility of a summary based on its generation probability conditioned on the source document, using a summarization model.

These methods are effective for detecting factual errors and inconsistencies in summaries. However, they mainly focus on whether a summary contradicts the original article. As a result, they do not directly address exaggerated expressions that emphasize specific aspects of the content. This limitation remains even when the summary is factually consistent with the source text.

2.2 Studies on Exaggerated Summaries

Compared to studies on factual consistency, relatively few studies have focused on exaggerated summaries. Iwamoto and Shimada [6] constructed a dataset of exaggerated summaries by rewriting normal summaries into exaggerated ones using LLMs. They also evaluated exaggeration detection models using this dataset. While this dataset is useful for analyzing exaggeration-related phenomena, the exaggerated summaries are artificially generated. Therefore, they may not sufficiently reflect exaggeration patterns observed in real news article summaries. In addition, the dataset is annotated with binary labels. This makes it difficult to represent differences in the degree of exaggeration.

To address these limitations, Iwamoto and Shimada [7] proposed a method for constructing a dataset of real news article–summary pairs annotated with continuous exaggeration scores. In this study, multiple article–summary pairs are compared using an LLM. The summaries are then ranked based on their relative levels of exaggeration. After that, continuous exaggeration scores are assigned to each summary by linearly mapping the ranking positions to a continuous scale. This framework allows exaggeration to be treated as a continuous quantity rather than a binary category. However, this approach assumes that exaggeration levels are uniformly distributed across the ranking. As a result, summaries located at intermediate positions in the ranking are assigned scores only based on their relative order. Semantic differences between summaries at neighboring ranks are not explicitly considered.

The ranking-based framework provides an important foundation for continuous exaggeration scoring. However, there is still room for improvement in how scores are assigned. This issue is especially important for summaries that fall between clearly non-exaggerated and clearly exaggerated cases. In this study, we build upon this prior work and investigate a scoring method that focuses on semantically meaningful intermediate points in the ranking. Our goal is to improve the interpretability of exaggeration score assignment.

3. Proposed Method

This section proposes a method for constructing an exaggeration score dataset that quantifies the degree of exaggeration in news article summaries as continuous values. The goal of the proposed method is to automatically assign exaggeration scores to a large set of article–summary pairs based on evaluations performed by an LLM.

The proposed framework consists of two stages. The first stage constructs an exaggeration ranking using LLM-based relative evaluation (Sorting). The second stage converts this ranking into continuous exaggeration scores (Scoring). While the Sorting stage follows an existing ranking-based framework [7], the Scoring stage is newly proposed in this study.

Before describing the overall framework, we first clarify the roles of absolute and relative evaluation in LLM-based exaggeration assessment.

3.1 Absolute and Relative Evaluation

There are two main approaches for evaluating the degree of exaggeration in news summaries using an LLM: absolute evaluation and relative evaluation.

In absolute evaluation, a single article–summary pair is given to the LLM, and the model directly outputs a numerical value that represents how exaggerated the summary is relative to the article. This approach is intuitive and easy to interpret. However, previous studies have reported that absolute evaluation often suffers from low consistency [8]. Even when the same input is provided repeatedly, the output score may vary across runs. This instability is mainly due to the fact that LLMs do not possess a clear internal scale and that their inference process is inherently probabilistic.

In contrast, relative evaluation presents two article–summary pairs to the LLM at the same time. The model is asked to compare them and decide which summary is more exaggerated, or which one is closer to a reference level. LLMs tend to produce more stable outputs in such comparative judgments than in absolute numerical judgments. However, relative evaluation only provides binary relationships between pairs and does not directly yield continuous exaggeration scores.

Table 1 summarizes the characteristics of absolute and relative evaluation. Based on this comparison, this study focuses on relative evaluation because of its higher consistency. To address the limitation that relative evaluation cannot directly output numerical scores, we adopt a framework that aggregates many relative comparison results and represents exaggeration as continuous values. Specifically, we first organize the relative order of exaggeration for all pairs and then construct continuous exaggeration scores based on this ordering.

Table 1 Absolute evaluation vs. Relative evaluation.

Method	Continuous Score	Consistency
Absolute	✓ Yes	✗ No
Relative	✗ No	✓ Yes

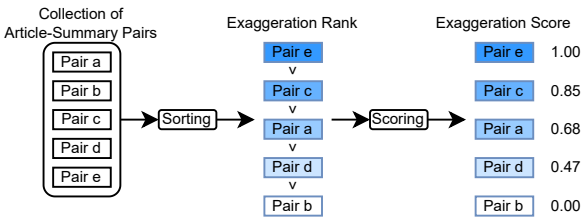

Fig. 2 Overview of the framework.

3.2 Overview of the Framework

Figure 2 shows an overview of the proposed framework for constructing an exaggeration score dataset. The framework consists of two stages: Sorting, which constructs an exaggeration ranking, and Scoring, which converts the ranking into continuous scores.

In the Sorting stage, a set of article–summary pairs is first prepared, as illustrated on the left side of Figure 2. An overview of the Sorting stage is illustrated in Figure 3. These pairs are ordered from the least exaggerated to the most exaggerated based on relative evaluations performed by an LLM. This stage follows the approach proposed by Iwamoto and Shimada [7], in which the LLM is used as a comparison function. Specifically, two article–summary pairs are presented to the model at a time. The model judges which summary is more exaggerated relative to its corresponding article, and the result of this judgment is used as the comparison outcome in the sorting procedure.

If all possible pairs were compared directly, the number of required comparisons would be on the order of N^2, where N is the number of pairs. This is not practical when N becomes large. Therefore, we adopt a comparison-based sorting algorithm. In this study, MergeSort is employed, which reduces the number of required comparisons to the order of $N \log N$. By combining MergeSort with LLM-based relative evaluation, we can construct an exaggeration ranking with a practical computational cost. The obtained ranking represents ordinal relationships among article–summary pairs in terms of their exaggeration degree.

In the Scoring stage, the obtained exaggeration ranking is used as input, and a continuous exaggeration score is assigned to each pair. While the ranking provides information about relative order, it does not indicate how large the differences in exaggeration are. Therefore, the Scoring stage aims to convert ordinal information into interpretable numerical scores while preserving the ranking order. This stage constitutes the main contribution of this study and is described in detail in the following sections.

3.3 Scoring: Converting the Ranking into Scores

This section describes a method for assigning continuous exaggeration scores to each article–summary pair, using the exaggeration ranking obtained in the Sorting stage as input.

The exaggeration ranking produced by Sorting represents the

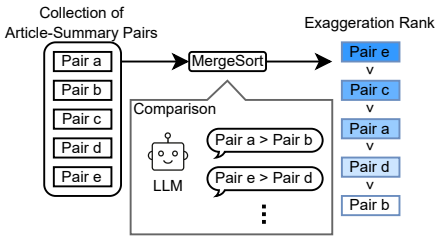

Fig. 3 Overview of the exaggeration ranking construction method.

relative order of pairs in terms of exaggeration. In other words, it guarantees which summary is more exaggerated than another. However, this information is purely ordinal. It does not indicate how large the difference in exaggeration is between two pairs. Nevertheless, in practical applications, it is often desirable that exaggeration scores have interpretable differences. For example, when the scores are used as target variables in regression-based learning, or when they are presented intuitively on social media platforms, simple rank information is not sufficient. In such cases, score differences are expected to reflect meaningful differences in exaggeration degree. Therefore, in the Scoring stage, it is necessary not only to preserve the ranking order but also to consider the interpretability of score differences.

The simplest approach to score assignment is to linearly map the ranking onto the interval $[0, 1]$. Specifically, a score of 0 is assigned to the least exaggerated pair, and a score of 1 is assigned to the most exaggerated pair. The remaining pairs are assigned evenly spaced values according to their rank positions. This method reflects the ranking order, such that a higher rank corresponds to a higher score. When a large collection of article–summary pairs is considered, the pairs located at both ends of the ranking can be regarded as approximations of the minimum and maximum observable exaggeration degrees. That is, the pairs ranked first and N-th can be interpreted as representing the lower and upper bounds of exaggeration in the data. Under this statistical assumption, assigning scores of 0 and 1 to the endpoints has a certain degree of validity.

However, linear assignment merely maps ordinal information onto an evenly spaced numerical scale. For example, a score of 0.5 simply represents the median rank and does not guarantee that the summary is semantically intermediate in terms of exaggeration. Similarly, the difference between scores 0.25 and 0.5 does not necessarily correspond to the same exaggeration difference as that between 0.5 and 0.75. This limitation can affect the interpretability of scores when they are used for learning or evaluation.

Therefore, in this study, we aim not only to preserve the ranking order but also to assign scores in a way that allows score differences to be interpreted as differences in exaggeration degree. To achieve this, we seek to construct exaggeration scores that can be understood as an interval scale rather than a purely ordinal scale.

3.3.1 Monotonicity and Equal-Interval Property

In this study, we assume that for each article–summary pair x, there exists an unobservable true exaggeration degree $e(x) \in [0, 1]$. Here, $e(x)$ is a latent variable that represents the strength of exaggeration in the summary relative to the article. Although

$e(x)$ cannot be directly observed, its relative order can be estimated through pairwise comparison. The exaggeration score $s(x) \in [0, 1]$ is defined as a mapping from the true exaggeration degree $e(x)$.

The minimum requirement for the score is monotonicity. That is, for any two pairs x_a and x_b, the score should preserve the order of the true exaggeration degrees, as formalized in Eq. 1. Linear assignment preserves the order of the ranking and therefore satisfies this requirement.

$$s(x_a) < s(x_b) \Leftrightarrow e(x_a) < e(x_b) \tag{1}$$

However, monotonicity alone does not guarantee a meaningful interpretation of score differences. For example, even when monotonicity holds, it is unclear whether a score difference of $s(x_b) - s(x_a) = 0.1$ corresponds to a negligible difference in exaggeration or to a non-negligible one. Similarly, a score of $s(x) = 0.5$ does not necessarily indicate that the exaggeration degree is semantically moderate. In addition, the difference between scores 0.25 and 0.5 is not guaranteed to correspond to the same exaggeration difference as that between 0.5 and 0.75. Thus, a score that satisfies only monotonicity is appropriate as an ordinal scale. However, score differences cannot be directly interpreted as differences in exaggeration degree.

To make scores more interpretable, this study introduces the equal-interval property, with the aim of constructing scores that can be understood as an interval scale. Specifically, the equal-interval property requires that a midpoint in score space corresponds to a semantic midpoint in exaggeration degree, as expressed in Eq. 2.

$$s(x_m) = \frac{s(x_a) + s(x_b)}{2} \Rightarrow e(x_m) \approx \frac{e(x_a) + e(x_b)}{2} \tag{2}$$

Intuitively, a score of 0.5 represents the semantic midpoint between 0 and 1, and a score of 0.25 represents the semantic midpoint between 0 and 0.5. If the equal-interval property is satisfied, score differences can be treated as differences in exaggeration degree. As a result, the suitability of the scores as target variables in learning tasks is also expected to improve.

However, because the target of evaluation is qualitative text, it is difficult to directly ask an LLM to evaluate quantitative concepts such as averages or numerical differences of exaggeration degrees. In the next section, we describe a method for identifying semantic midpoints based solely on relative comparison, without relying on numerical comparison.

3.3.2 Identification of Semantic Midpoints Based on Relative Comparison

This section describes a method for identifying semantic midpoints based solely on relative comparison, under the assumption that the set of article–summary pairs has already been sorted in increasing order of exaggeration by the Sorting stage.

As an initial condition, we assume that a score of 0 is assigned to the lowest-ranked pair in the exaggeration ranking, and a score of 1 is assigned to the highest-ranked pair. This assumption is based on the statistical consideration that, when a large collection of article–summary pairs is used, the pairs located at both ends of the ranking can be regarded as approximations of the observable lower and upper bounds of exaggeration. Under this initial condition, only the two endpoint pairs have known scores, while the scores of pairs located between them remain undetermined.

At this stage, the search interval for midpoint identification, denoted as $[L, R]$, is defined as follows. The index L corresponds to the position immediately after the lower-exaggeration anchor, and the index R corresponds to the position immediately before the higher-exaggeration anchor. In other words, the search interval contains pairs whose scores are unknown and is bounded by two anchor pairs with known scores. Let x_L denote the lower-exaggeration endpoint and x_R denote the higher-exaggeration endpoint. Our goal is to identify, within this interval, the pair that corresponds to the semantic midpoint between x_L and x_R.

For a candidate pair x_M within the search interval, the proposed method asks the LLM to perform a relative comparison. Specifically, the LLM is asked whether x_M is closer in exaggeration degree to x_L or to x_R. By using the result of this proximity-based comparison, the midpoint identification process can be interpreted as a search problem. If x_M is judged to be closer to the lower-exaggeration endpoint x_L, the semantic midpoint is considered to lie on the higher-exaggeration side. In this case, the search interval is updated to $[M + 1, R]$. Conversely, if x_M is judged to be closer to the higher-exaggeration endpoint x_R, the search interval is updated to $[L, M - 1]$. By repeatedly updating the search interval in this manner, the position where the distances to the two endpoints are balanced can be gradually narrowed down. This procedure is analogous to binary search in index space, under the assumption that exaggeration increases monotonically along the ranking.

Figure 4 illustrates a concrete example of this midpoint identification process. We assume that the lowest-ranked pair has already been assigned a score of 0 and the highest-ranked pair has been assigned a score of 1. The undetermined interval $[L, R]$ between these two anchors is the target of the search. In the first step, a candidate index M located near the center of the interval is selected. The LLM then judges whether the pair at index M is closer in exaggeration to the pair at index L or to the pair at index R. For explanation purposes, we assume the existence of a true exaggeration degree $e(x)$. In the example shown in the figure, the true exaggeration degrees are assumed to be $e(a) = 0$ for the pair at L, $e(k) = 1$ for the pair at R, and $e(f) = 0.67$ for the candidate pair at M. Since $e(f)$ is closer to $e(k)$ than to $e(a)$, it is natural to judge that the candidate pair M is closer to the higher-exaggeration side. The relative comparison performed by the LLM is expected to reflect this relationship and output that M is closer to x_R. Based on this result, the semantic midpoint is considered to lie toward the lower-exaggeration side, and the search interval is updated to $[L, M - 1]$. By repeating this update process based on proximity judgments, the search interval is gradually narrowed to the position that best balances the distances to the two endpoints. When the search interval converges to a single pair, that pair is interpreted as the one closest to the semantic midpoint between scores 0 and 1. In the example, a score of 0.5 is assigned to pair d, in Figure 4.

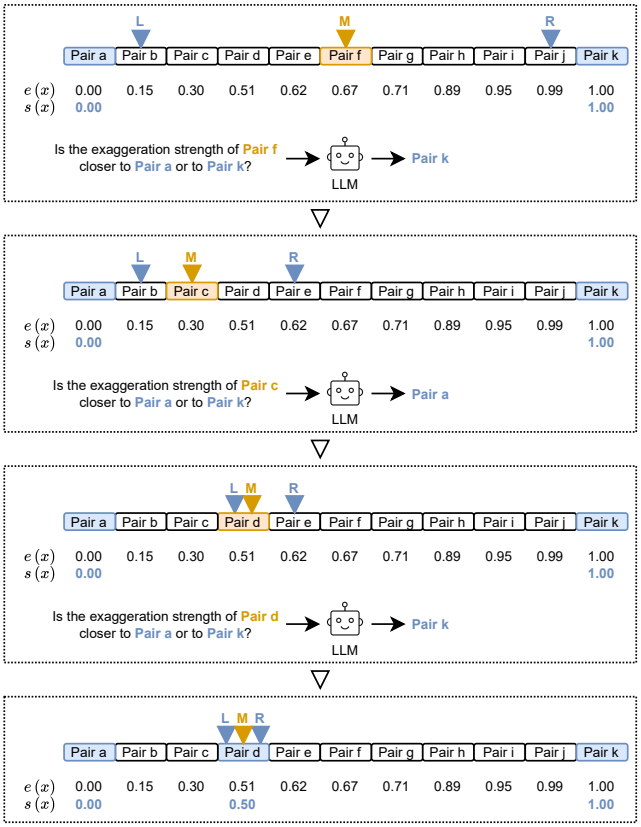

Fig. 4 Example of semantic midpoint identification based on relative comparison.

3.3.3 Recursive Application of Midpoint Identification and Construction of Anchor Points

This section describes a method for constructing multiple anchor points by recursively applying the midpoint identification procedure, thereby dividing the score scale in a stepwise manner.

Specifically, in addition to the anchor points corresponding to scores 0 and 1, a score of 0.5 is assigned to the semantic midpoint identified by the procedure described in Section 3.3.2. Next, the intervals $[0, 0.5]$ and $[0.5, 1]$ are regarded as independent search intervals, and the same midpoint identification procedure is applied to each interval. As a result, pairs corresponding to the semantic midpoints of these intervals are identified, and scores of 0.25 and 0.75 are assigned, respectively. By recursively applying the midpoint identification in this manner, the score scale is gradually divided, and multiple anchor points such as 0, 0.25, 0.5, 0.75, and 1 can be obtained. These anchor points are not determined by simple uniform index partitioning. Instead, their semantic midpoints are confirmed through relative comparison.

The level of detail in the constructed anchor set involves a trade-off between computational cost and precision. Since midpoint identification requires relative comparisons performed by an LLM, the number of comparisons increases as the recursion depth becomes larger. Therefore, in the proposed method, the recursive application of midpoint identification is terminated after reaching a predefined depth, once a sufficient number of anchor points has been obtained.

Figure 5 illustrates a concrete example of the process in which the anchor set is constructed through recursive midpoint identification. At recursion depth 0, which corresponds to the initial

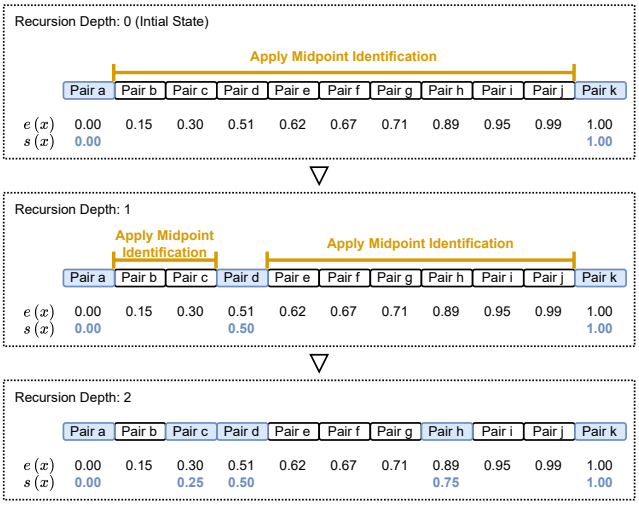

Fig. 5 Example of anchor point construction through recursive midpoint identification.

state, scores of 0 and 1 are assigned to the lowest-ranked and highest-ranked pairs in the exaggeration ranking, respectively. At recursion depth 1, midpoint identification is applied to the undetermined interval between these two anchors, and the pair corresponding to the semantic midpoint between scores 0 and 1 is identified and assigned a score of 0.5. At recursion depth 2, the intervals $[0, 0.5]$ and $[0.5, 1]$ are treated as independent search intervals. By applying the same midpoint identification procedure to each interval, anchor points corresponding to scores 0.25 and 0.75 are identified.

3.3.4 Computing Continuous Scores by Interpolation between Anchor Points

This section describes an interpolation method for assigning continuous exaggeration scores to all article–summary pairs based on the constructed set of anchor points.

In the proposed method, PCHIP (Piecewise Cubic Hermite Interpolating Polynomial) [9] is adopted as a technique for smoothly interpolating the entire score scale using the anchor points as reference points. PCHIP is a method based on piecewise cubic Hermite interpolation, which passes through all given data points while preserving monotonicity. In addition, it is known to be less prone to excessive oscillation, which often occurs in higher-order polynomial interpolation. These properties are useful for smoothly representing the relationship between the ranking and the scores, while faithfully preserving the anchor points obtained through midpoint identification.

Specifically, the index of the exaggeration ranking is used as the horizontal axis, and the exaggeration scores assigned to the anchor points are used as the vertical axis. By interpolating between adjacent anchor points using PCHIP, a continuous score function over the entire ranking is constructed. Using this function, a continuous exaggeration score $s(x) \in [0, 1]$ can be assigned to every article–summary pair in a consistent manner.

The scores obtained through this procedure guarantee monotonicity, as the order of the ranking is preserved. At the same time, the anchor points retain semantic interpretations based on equal partitioning, which was introduced through relative midpoint identification.

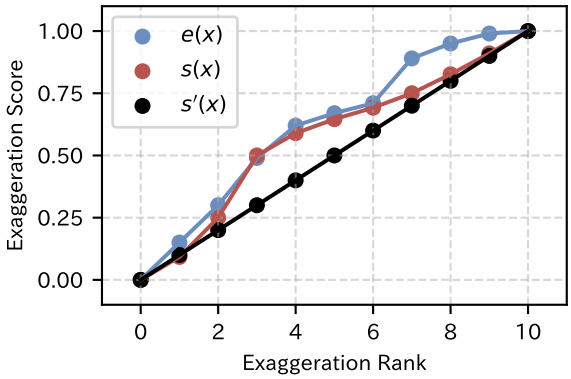

Fig. 6 Comparison of scores derived from the exaggeration ranking.

Figure 6 illustrates a concrete example of scores derived from an exaggeration ranking. In the figure, the blue points represent the true exaggeration degree $e(x)$, the red points represent the exaggeration scores $s(x)$ obtained by the proposed method, and the black points represent scores $s'(x)$ obtained by simple linear assignment based on the ranking. Since the true exaggeration degree $e(x)$ is unobservable, this figure is a schematic example constructed using hypothetical values for explanatory purposes. Although linear assignment guarantees monotonicity, it uniformly converts rank differences into numerical values. As a result, discrepancies may arise when the true exaggeration degree is distributed in a nonlinear manner. In contrast, the scores obtained by the proposed method introduce equal partitioning through midpoint identification based on relative comparison. This enables the construction of scores that preserve monotonicity while more closely reflecting the true exaggeration degree.

3.4 Enhancing Robustness of the Proposed Method

LLMs are based on probabilistic generation processes. Therefore, even when the same prompt is given, the judgment results may vary across different executions. Such variability in the outputs can influence both the Sorting stage, which relies on relative comparisons, and the Scoring stage, which converts rankings into continuous scores. To address this issue, the proposed method improves robustness by aggregating the results of multiple executions in both stages.

First, in the Sorting stage, the initial order of article–summary pairs is randomly shuffled, and the construction of the exaggeration ranking using MergeSort is performed multiple times. For each execution, a ranking of the pairs is obtained based on relative comparisons by the LLM. After performing multiple executions, the average rank of each pair is computed across all rankings. The pairs are then re-sorted according to these average ranks to form the final exaggeration ranking. This procedure is expected to reduce biases that depend on a specific initial ordering or a single execution result.

Next, in the Scoring stage, the averaged exaggeration ranking obtained in the Sorting stage is used as input. The entire scoring procedure, which consists of midpoint identification, construction of the anchor point set, and computation of continuous scores by interpolation, is executed multiple times. For each article–summary pair, the scores obtained from these multiple executions are averaged. The resulting average value is used as the final exaggeration score.

By introducing ensemble processing in both the Sorting and Scoring stages, the proposed method suppresses errors caused by variability in the probabilistic outputs of the LLM. As a result, the stability of both the exaggeration ranking and the assigned scores is improved.

4. Experiments

In this section, the proposed method is applied to a real-world news article corpus to construct an exaggeration score dataset, and its properties are examined through experimental evaluation. Since the exaggeration scores proposed in this study are constructed based on relative evaluations performed by an LLM, it is important to verify whether the assigned scores are obtained in a stable manner without strong dependence on execution conditions, and whether they appropriately reflect the degree of exaggeration.

From this perspective, we focus on evaluating the stability of the results produced in the Scoring stage (Section 3.3), which converts the exaggeration ranking into continuous scores. In addition, we examine the validity of the assigned scores using artificially generated exaggerated summaries. In particular, the experiments investigate how variability arising from the probabilistic generation process of the LLM affects fluctuations in the assigned scores, and whether the scores appropriately capture the presence and strength of exaggeration.

4.1 Experimental Setup

In this experiment, we use 1,000 article–summary pairs selected from the Newsroom corpus [10], which contains a large collection of real-world news articles and their corresponding summaries. The proposed method is applied to this corpus to construct an exaggeration score dataset.

In the Scoring stage, the recursion depth for constructing the anchor point set is fixed to 10. As the LLMs used for relative evaluation, `Mistral-7B-Instruct-v0.3`[*1] is employed in the Sorting stage, and `gpt-oss-20b`[*2] is used in the Scoring stage. For both models, the temperature parameter during generation is fixed to 0.8.

In the Sorting stage, the LLM is given two article–summary pairs at a time and is asked to judge which summary is more exaggerated relative to its corresponding article.

In the Scoring stage, relative evaluation is used for midpoint identification. The LLM is presented with one target article–summary pair and two reference pairs. It is then asked to determine whether the exaggeration level of the target summary is closer to that of the first reference pair or the second, relative to each original article. The prompt explicitly instructs the model to ignore content similarity and to base the judgment solely on the degree of exaggeration.

The proposed method is applied to the corpus to construct the exaggeration score dataset used in the subsequent experiments.

[*1] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`
[*2] `https://platform.openai.com/docs/models/gpt-oss-20b`

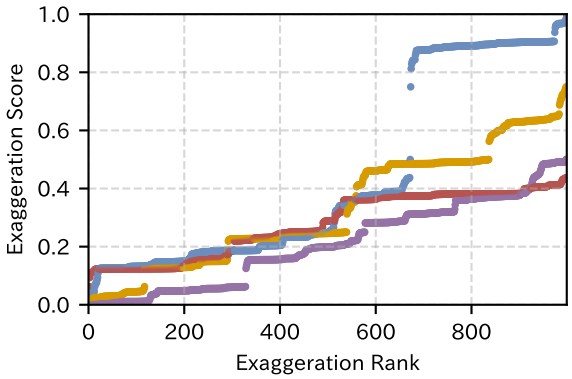

Fig. 7 Relationship between exaggeration ranking and exaggeration scores across four score assignment runs.

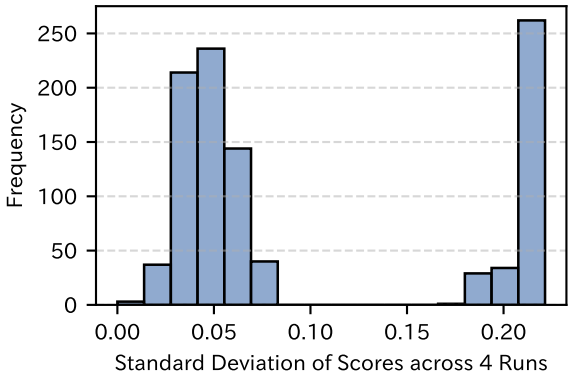

Fig. 8 Distribution of score standard deviations across four score assignment runs.

4.2 Stability of Score Assignment Results

This section examines how stable the score assignment results are with respect to differences across multiple executions of the scoring procedure. In the proposed method, the final exaggeration scores are obtained through the score assignment procedure applied to a fixed ranking. Since the score assignment procedure relies on relative comparisons performed by an LLM and subsequent interpolation, it is important to evaluate how stable the assigned scores are across different execution runs. Therefore, this section focuses on examining the stability of the score assignment process itself. To isolate the effect of score assignment from fluctuations in the sorting results, we fix the ranking and execute the scoring procedure multiple times. Specifically, following the aggregation method described in Section 3.4, a single ranking is constructed by aggregating four sorting results. This aggregated ranking is then used as the fixed input, and the scoring procedure is executed four times. In each execution, the seed value of the LLM used for score assignment is changed, and midpoint identification as well as the subsequent interpolation process are performed independently.

First, to qualitatively examine the stability of the score assignment results, we visualize the relationship between the ranking and the assigned scores. Figure 7 shows the correspondence between the exaggeration ranking on the horizontal axis and the exaggeration scores on the vertical axis for the four executions. Each plot in the figure corresponds to one execution with a different seed value, and the color indicates the execution run. From the figure, it can be observed that from the lower ranks to approximately the middle ranks around rank 600, similar score assignment patterns are obtained across all four executions. This indicates that the correspondence between the ranking and the scores is stable in this range. In contrast, in the higher-ranked region, some discrepancies can be observed between certain executions and the others. Nevertheless, a common trend of rapidly increasing scores in the higher-ranked region is consistently observed across all executions. These observations suggest that the score assignment process produces generally stable results.

Next, we quantitatively evaluate the variability of the score assignment results. For all 1,000 article–summary pairs, the scores obtained in the four executions are collected, and the standard deviation of the scores is computed for each pair. As a result, the mean of the standard deviations is 0.1008. Considering that the exaggeration scores range from 0 to 1, this value corresponds to approximately 10% of the entire score range. This indicates that the score assignment results exhibit a certain level of stability across different executions.

Figure 8 shows the distribution of the standard deviations of the scores for all pairs obtained from the four executions. The horizontal axis represents the standard deviation of the scores across the four runs, and the vertical axis represents the number of corresponding pairs. The figure shows that the distribution has concentrations around 0.05 and around 0.20. Based on the results shown in Figure 7, the pairs with relatively large standard deviations are considered to be influenced by variations in score assignment that occur in some executions within the higher-ranked region. However, the standard deviations are small for most pairs, and cases in which score assignment becomes unstable are limited.

From these results, we conclude that under the condition where the ranking is fixed, the score assignment process in the proposed method exhibits a sufficient level of stability with respect to differences across execution runs.

4.3 Validity of Scores Using Artificially Exaggerated Summaries

This section evaluates whether the scores assigned by the proposed method appropriately reflect the degree of exaggeration contained in summaries. Specifically, we examine the relationship between the assigned scores and binary labels indicating the presence or absence of exaggeration, using artificially generated exaggerated summaries.

In this experiment, we generate artificial exaggerated summaries based on the CNN/DailyMail dataset [11], following the dataset construction method proposed by Iwamoto and Shimada [6]. For each original summary, GPT-4o[*3] is used to rewrite the summary by adding only one exaggerated expression, while explicitly avoiding the introduction of new factual errors such as contradictions or incorrect numerical information. As a result, the generated exaggerated summaries differ from the original ones only in the presence of an exaggerated expression, allowing controlled evaluation of exaggeration.

Through this process, we construct a set of 1,000 article–summary pairs, consisting of 500 original summaries with-

[*3] https://platform.openai.com/docs/models/gpt-4o

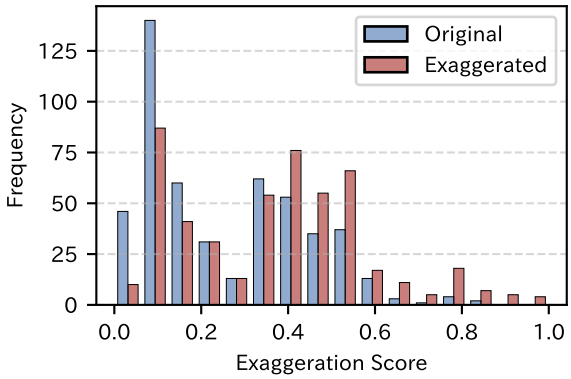

Fig. 9 Distribution of exaggeration scores for original and artificially exaggerated summaries.

out exaggeration and 500 artificially generated exaggerated summaries. The original summaries and exaggerated summaries are created from different source articles, so that no articles overlap between the two groups. This design avoids cases in which judgments become easier due to content similarity arising from the same original article, and ensures that the evaluation focuses on exaggeration itself.

The proposed method is then applied to these 1,000 pairs to assign exaggeration scores to all summaries. As a result, the average score of the original summaries is 0.2576, while the average score of the artificially exaggerated summaries is 0.3698. Thus, the exaggerated summaries consistently receive higher scores than the non-exaggerated summaries. This result indicates that the scores assigned by the proposed method appropriately reflect the presence of exaggeration.

Figure 9 shows the distribution of exaggeration scores for the original summaries and the artificially generated exaggerated summaries. The horizontal axis represents the exaggeration scores assigned by the proposed method, and the vertical axis represents the number of corresponding summaries. In the figure, the blue histogram corresponds to the original summaries without exaggeration, and the red histogram corresponds to the artificially exaggerated summaries. It can be observed that the proportion of exaggerated summaries increases as the score becomes higher, while original summaries are more densely distributed in the lower score region.

From these results, we conclude that the proposed scores can smoothly distinguish between exaggerated and non-exaggerated summaries, and that they appropriately reflect the degree of exaggeration.

5. Conclusion

In this study, we proposed a method for constructing an exaggeration score dataset that represents the degree of exaggeration in news article summaries as continuous numerical values. This work is motivated by the observation that exaggerated expressions in summaries can influence readers' understanding of information and potentially affect the formation of public opinion. To address this issue, we designed a two-stage framework consisting of Sorting and Scoring.

However, this study has several limitations. First, the semantic validity of the exaggeration scores has been evaluated mainly using artificially generated exaggerated summaries and illustrative analysis. Because the perception of exaggeration depends on readers' background knowledge and values, future work should incorporate human annotations and more diverse evaluations to further examine the interpretability and validity of the scores.

Second, the proposed method constructs exaggeration scores based on relative evaluations by an LLM. As a result, the outcomes may still be influenced by the choice of the LLM and the prompt design. Further investigation is required to determine whether similar tendencies are observed across different settings.

As future work, we plan to use the constructed exaggeration score dataset to train models that automatically estimate the degree of exaggeration in summaries. Such models would enable automatic estimation of exaggeration levels for large volumes of news article summaries in practical settings, and could be applied to applications such as providing warnings or alerts to readers.

References

[1] Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A. and Bavel, J. J. V.: Emotion shapes the diffusion of moralized content in social networks, *Proceedings of the National Academy of Sciences*, Vol. 114, No. 28, pp. 7313–7318 (online), DOI: 10.1073/pnas.1618923114 (2017).

[2] Vosoughi, S., Roy, D. and Aral, S.: The spread of true and false news online, *Science*, Vol. 359, No. 6380, pp. 1146–1151 (online), DOI: 10.1126/science.aap9559 (2018).

[3] Kryscinski, W., McCann, B., Xiong, C. and Socher, R.: Evaluating the Factual Consistency of Abstractive Text Summarization, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Webber, B., Cohn, T., He, Y. and Liu, Y., eds.), Online, Association for Computational Linguistics, pp. 9332–9346 (online), DOI: 10.18653/v1/2020.emnlp-main.750 (2020).

[4] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y.: BERTScore: Evaluating Text Generation with BERT, *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, (online), available from ⟨https://openreview.net/forum?id=SkeHuCVFDr⟩ (2020).

[5] Yuan, W., Neubig, G. and Liu, P.: Bartscore: Evaluating generated text as text generation, *Advances in neural information processing systems*, Vol. 34, pp. 27263–27277 (2021).

[6] Iwamoto, K. and Shimada, K.: Dataset Construction and Verification for Detecting Factual Inconsistency in Japanese Summarization, *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 243–248 (online), DOI: 10.1109/IIAI-AAI63651.2024.00054 (2024).

[7] Iwamoto, K. and Shimada, K.: Exaggeration Scoring of News Summaries through LLM-based Relative Judgments, *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, Hanoi, Vietnam (2025).

[8] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T. and Sui, Z.: Large Language Models are not Fair Evaluators, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Ku, L.-W., Martins, A. and Srikumar, V., eds.), Bangkok, Thailand, Association for Computational Linguistics, pp. 9440–9450 (online), DOI: 10.18653/v1/2024.acl-long.511 (2024).

[9] Fritsch, F. N. and Butland, J.: A Method for Constructing Local Monotone Piecewise Cubic Interpolants, *SIAM Journal on Scientific and Statistical Computing*, Vol. 5, No. 2, pp. 300–304 (online), DOI: 10.1137/0905021 (1984).

[10] Grusky, M., Naaman, M. and Artzi, Y.: NEWSROOM: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 708–719 (online), available from ⟨http://aclweb.org/anthology/N18-1065⟩ (2018).

[11] Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç. and Xiang, B.: Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning* (Riezler, S. and Goldberg, Y., eds.), Berlin, Germany, Association for Computational Linguistics, pp. 280–290 (online), DOI: 10.18653/v1/K16-1028 (2016).