# Exaggeration Scoring of News Summaries through LLM-based Relative Judgments

**Keisuke Iwamoto**
Department of Creative Informatics
Kyushu Institute of Technology
Fukuoka, Japan
iwamoto.keisuke629@mail.kyutech.jp

**Kazutaka Shimada**
Department of Artificial Intelligence
Kyushu Institute of Technology
Fukuoka, Japan
shimada@ai.kyutech.ac.jp

## Abstract

Exaggerated summaries in news articles mislead readers and cause the spread of misinformation, especially on social media, where short and eye-catching content is common. Previous studies have tried to detect exaggeration using classification-based methods. However, they usually use binary labels and do not consider different levels of exaggeration. In this study, we propose a ranking-based method to create a dataset with continuous exaggeration scores for news summaries. We use a large language model (LLM) to compare how exaggerated different article-summary pairs are. By running MergeSort multiple times using the LLM as a comparison function, we can rank the summaries based on their exaggeration. Then, we combine the results from all the sorting runs to assign stable and reliable exaggeration scores. Our experiments show that these scores are consistent across sorting trials, match human intuition well and are effective in identifying artificially exaggerated summaries generated by GPT-4o. These results suggest that our LLM-based ranking approach can provide a solid basis for measuring exaggeration levels in text summaries. This can help improve the training and evaluation of models that detect exaggeration in a more detailed and accurate way.

## 1 Introduction

In recent years, social networking services (SNS) such as X and Facebook have become important platforms where many people share news articles and their opinions. On these platforms, users often read and spread information in short forms, like headlines or summaries. These short versions are sometimes automatically generated or made by users and other parties. Because of the short character limits and fast-paced nature of SNS, these summaries can spread more quickly than full articles.

However, this kind of information sharing brings a serious problem. Some summaries or headlines are exaggerated. This means that they emphasize or highlight certain parts of the original article too much and give stronger impressions than the original text. Even though these summaries are not always factually incorrect, they often use emotional or exciting expressions and focus on small parts of the article. As a result, they may cause readers to misunderstand the original content, and this can lead to biased opinions or wrong public understanding. For example, here are two summaries of the same article about university baseball rankings:

**Normal** "The article introduces the predicted rankings of college baseball teams for the upcoming season."

**Exaggerated** "Top university in SHOCK as baseball team DROPS out of championship race!"

Both summaries talk about the same topic, but the exaggerated one uses strong words to make it sound more dramatic. This could give readers the wrong impression, as if something very serious or shocking happened.

Exaggerated summaries are especially problematic because they are more likely to go viral than accurate summaries. Previous studies on information sharing show that emotional or sensational content spreads more widely on SNS than neutral content (Brady et al., 2017; Vosoughi et al., 2018). Moreover, SNS users usually have a short attention span. As a result, misleading summaries quickly gain popularity and influence people's opinions on serious topics.

To address this problem, we aim to develop a method that can indicate how exaggerated a summary is in a way that is both quick and easy to understand. Since people do not have time to read long explanations on SNS, we suggest using a numeric exaggeration score as the first step to alert users. This score would help users quickly know

whether they should be careful about the content they are reading.

This kind of score has several benefits:

- It is easier to understand than a long written explanation.
- Users can easily compare different summaries.
- It can be used in systems like automatic moderation or content ranking.

Although written explanations can give more details, they are often too slow for social media. A number-based score is faster and more practical, and it helps balance freedom of expression with the need to reduce the spread of misleading content.

In addition, the score must be automatically generated, because it is impossible to check every summary by hand due to the large amount of content. To achieve this, we created a large-scale dataset with article-summary pairs. Each pair is given a numeric exaggeration score that shows how much the summary exaggerates the original article. This dataset will help us train machine learning models in the future to automatically predict exaggeration scores.

In this study, we propose a method to build a large dataset of news article-summary pairs with exaggeration scores. We also explore how to use large language models (LLMs) to make these exaggeration annotations in a stable and consistent way. Our final goal is to support the development of machine learning models that can detect and score exaggeration in real-world summaries. We hope that our work can help to promote more accurate and responsible information sharing on social media.

Figure 1 shows the overall process of our method. First, we collect article-summary pairs from existing datasets. Next, an LLM compares pairs of summaries to see which one is more exaggerated. We run a sorting algorithm (MergeSort) several times with different orders and take the average results to get stable exaggeration scores for each summary. These scores are used to create a dataset, which can be used to train exaggeration detection models. The top part of the figure shows the main focus on this paper: building a reliable exaggeration score dataset using pairwise comparisons by a language model.

Main contributions of this work are:

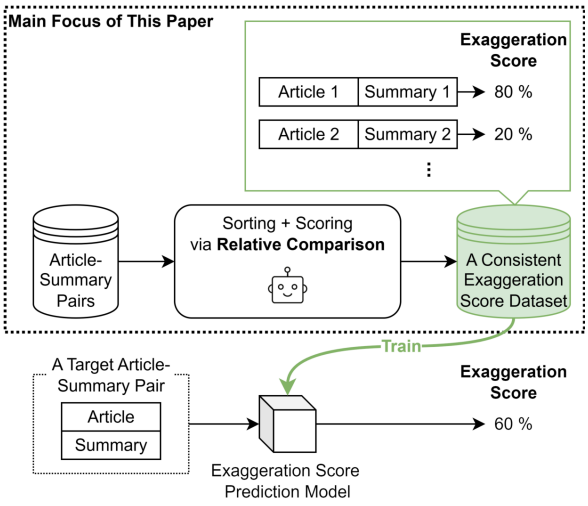


Figure 1: Overview of our approach. LLM-based relative comparisons are used to score summaries, which are then used to train exaggeration detection models. The upper part highlights the focus of this paper.

- We introduce a new task of assigning exaggeration scores to summaries of news articles. This enables a more fine-grained measurement compared to binary classification.
- We propose a two-step method using pairwise comparison by LLMs and sorting algorithms to generate stable exaggeration scores.
- We build a large dataset by applying our method to many real-world news article-summary pairs.
- We perform detailed evaluations using both statistical methods and human feedback to confirm the stability and usefulness of the scores.

## 2 Related Work

### 2.1 Factual Consistency in Summarization

Many researchers have studied how to check if generated summaries are factually correct. Some methods try to find if the summary has wrong or made-up information that is not in the original text. Well-known methods include:

**FactCC (Kryscinski et al., 2020)** A model that classifies whether the summary and the source document match in meaning.

**BERTScore (Zhang et al., 2020)** A method that compares the summary and the source using embeddings from a pre-trained language model.

**BARTScore (Yuan et al., 2021)** A score based on how likely a summary is, given the source text, using the BART model.

These methods are good for checking factual correctness. However, they don't check if the summary is emotionally exaggerated or too dramatic. Even if a summary is factually correct, it can still give a wrong impression because of exaggerated expressions.

### 2.2 Exaggerated Summary Detection

Only a few studies focus on finding exaggeration in text. A dataset where a language model rewrote normal summaries in an exaggerated way was created (Iwamoto and Shimada, 2024). This is useful for exaggeration-related tasks. However, there are some problems:

- The exaggerated summaries are artificial and may not show the kind of exaggeration we often see in real social media or news.
- The dataset only gives binary labels (exaggerated or not), so it cannot show how much the summary is exaggerated.

In our study, we try to give continuous exaggeration scores instead of just binary labels. This allows us to do more detailed analysis and supports tasks like ranking or filtering summaries based on exaggeration levels.

### 2.3 LLM-based Evaluation Paradigms

Recently, people have started using large language models (LLMs) to evaluate summaries or answers. LLMs are good at comparing two texts and deciding which one is better. However, they are not good at giving consistent scores for single texts, because their scoring is not stable (Wang et al., 2024; Zheng et al., 2023).

Based on this, we use a pairwise ranking method to build our exaggeration score dataset. Instead of giving scores directly, we compare pairs of summaries and then turn the results into numbers using ranking and normalization. This helps us get more stable and reliable scores.

## 3 Proposed Method

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in understanding complex natural language, including the ability to perform contextual reasoning, semantic comparison, and pragmatic inference across textual inputs. These models have shown success in various evaluative tasks such as factuality judgment, content ranking, and summarization evaluation, often rivaling human-level performance in pairwise decision-making scenarios (Wang et al., 2024; Zheng et al., 2023).

In particular, the task of assessing how much a summary exaggerates the original article requires not just surface-level textual matching. However, it also involves deep semantic alignment, nuanced tone detection, and discourse-level interpretation. These are precisely the kinds of tasks that LLMs are well-suited for.

Thus, our method leverages LLMs as core evaluators to estimate exaggeration levels, based on their strong abilities to:

- Capture subtle differences in tone, emphasis, or framing between a summary and its source article.
- Understand the broader context and intent behind a summary, beyond factual correctness.
- Make robust comparative judgments between multiple summaries, even when exaggeration is implicit or stylistic rather than explicitly false.

We therefore adopt LLMs not only as a technical tool, but as an essential enabler for building a high-fidelity exaggeration scoring system grounded in semantic understanding.

### 3.1 Evaluation Strategies: Absolute vs. Relative

A straightforward approach to exaggeration scoring is absolute evaluation, in which an LLM receives a single article-summary pair and directly outputs a score. While this method is simple to implement, prior work (Wang et al., 2024) has shown that LLMs tend to produce unstable or inconsistent scores for the same inputs. This instability is attributed to the absence of an internal absolute standard for scoring within LLMs and their inherent non-determinism. As a result, direct scoring suffers from poor reproducibility and limited reliability.

In contrast, relative evaluation prompts an LLM to compare two different article-summary pairs and decide which summary is more exaggerated. LLMs

Table 1: Comparison of Absolute vs. Relative Evaluation Methods.

| Evaluation Type | Scalar Score | Consistency |
|---|---|---|
| Absolute | ✓ Yes | ✗ No |
| Relative | ✗ No | ✓ Yes |

are generally more consistent and accurate when performing such pairwise comparisons, as they rely on relative distinctions rather than absolute criteria. However, the limitation of this approach is that it only yields binary judgments (e.g., "A is more exaggerated than B") and does not produce scalar scores directly.

The trade-offs between these two approaches are summarized in Table 1. To overcome the weaknesses of each method, we propose a hybrid framework that utilizes the consistency of relative evaluation to produce a large-scale dataset with high-quality exaggeration rankings, from which scalar scores are derived.

### 3.2 Overall Framework

Figure 2 shows the overall workflow of our exaggeration scoring method. In this paper, we focus on the first part of Figure 1 (in Section 1): making a dataset with exaggeration scores.

First, we collect article-summary pairs, such as Pair A, Pair B, and so on (shown on the left side of the figure). Then, we use an LLM to compare the pairs and decide which one is more exaggerated. For example, the model might say "Pair A is more exaggerated than Pair B" or "Pair E is more exaggerated than Pair D."

Next, we use a sorting algorithm to combine the comparison results and rank the pairs by exaggeration level (as shown on the right side of the figure). Finally, we convert the ranks into scores between 0 and 1. These scores make up our exaggeration score dataset.

### 3.3 Score Derivation via LLM-based Sorting

To generate scores from pairwise comparisons, we formulate the problem as a ranking task: Given $N$ article-summary pairs, we aim to sort them in increasing order of exaggeration. We apply a modified version of the MergeSort algorithm, where the LLM acts as the comparator function. This reduces the number of required comparisons from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$, making it feasible to process thousands of samples.

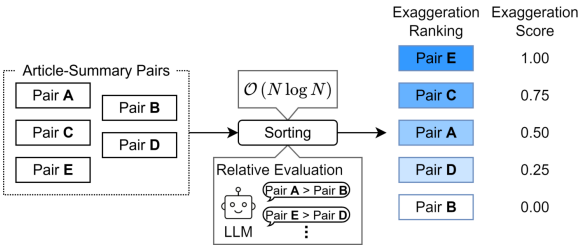


Figure 2: Overview of the proposed exaggeration scoring framework. Article-summary pairs are compared using LLMs, sorted by exaggeration level, and converted into scalar scores.

After sorting, we assign a score of 0 to the least exaggerated summary and a score of 1 to the most exaggerated one. The remaining summaries are linearly interpolated based on their relative position in the sorted list.

This linear normalization assumes that in a sufficiently large and diverse set of samples, the summary ranked at the very bottom (i.e., least exaggerated) can be treated as representing a near-zero exaggeration level in an absolute sense. Likewise, the summary ranked at the top (i.e., most exaggerated) can be considered to represent a maximally exaggerated instance, even without a formal definition of exaggeration intensity.

This assumption is supported by a statistical perspective: If the number of article-summary pairs $N$ is large (e.g., in the thousands), then the items at the extrema of the ranking (rank 1 and rank $N$) are likely to approximate the empirical minimum and maximum of exaggeration observable in real-world data. As such, mapping these ranks to the endpoints of the $[0, 1]$ scale provides a practical and interpretable score space, which reflects the relative exaggeration intensity in a pseudo-absolute fashion. This approach enables consistent score assignment even though the underlying comparisons are pairwise and relative in nature.

### 3.4 Enhancing Robustness

Since LLM outputs can be non-deterministic and sometimes biased due to input formatting, we introduce two techniques to enhance the robustness and reliability of the sorting results.

#### 3.4.1 Bidirectional Prompting

Previous work (Wang et al., 2024) has shown that LLMs may exhibit position bias in pairwise evaluations—that is, the summary shown earlier in the prompt may be judged differently than if shown second. This bias becomes particularly problematic

when comparing pairs whose exaggeration levels are close, as the model may be influenced more by order than content.

To address this, we perform comparisons in both directions: for a given pair (A, B), we prompt the LLM twice, once with A before B and once with B before A. If the results are consistent (e.g., A is judged more exaggerated than B in both cases), we apply the result to the sorting process. If the results conflict, we assume the difference is not clear and no reordering is performed. This filtering reduces the risk of introducing noise into the global ranking.

#### 3.4.2 Ensemble Sorting with Permutation Averaging

LLMs are inherently stochastic; even with fixed prompts, the outputs may vary across runs. To mitigate this, we apply the MergeSort-based ranking process multiple times with different randomly shuffled initial orders. For each run, we compute the full ranking of the dataset. Then, we aggregate these rankings by computing the average rank for each item and use this as the final basis for score assignment.

This ensemble approach reduces variance and increases stability. For example, if a pair consistently appears near the top of the ranking across runs, we can confidently assign it a low exaggeration score. Conversely, if its position fluctuates, it likely indicates ambiguity, which is reflected in its intermediate score.

## 4 Experiments

### 4.1 Experimental Settings

To evaluate how reliable and valid our proposed exaggeration scoring method is, we created a dataset of 1,000 article-summary pairs from the Newsroom corpus (Grusky et al., 2018), using the full scoring process described in Section 3.

#### 4.1.1 Dataset

We used the Newsroom dataset, which is a large-scale summarization corpus that contains article-summary pairs from various fields such as politics, science, and lifestyle. From this dataset, we randomly selected 1,000 article-summary pairs. To maintain good quality, we removed duplicates and extremely short summaries. As a result, the dataset includes diverse language expressions and keeps a certain level of summary quality.

#### 4.1.2 LLM Configuration

We used `Mistral-7B-Instruct-v0.3`[1] as the LLM for all pairwise comparison tasks in the sorting stage. The model was run in a local inference environment, which allowed us to fully control the temperature setting, prompt format, and output format. In our experiments, we set the temperature to 0.8.

In each comparison, the model was asked to evaluate two article-summary pairs and assign exaggeration scores from 0 to 3. It also had to give a reason for its decision. The prompt was designed so that the model would first explain its reasoning and then output the scores in JSON format. This approach follows previous research (Zheng et al., 2023), which shows that generating reasoning before giving scores, a method called Chain-of-Thought prompting, leads to more accurate results.

Importantly, although the model was not directly instructed to compare the two pairs against each other, it was asked to score both in a single prompt. This setting naturally encourages the model to indirectly compare the two pairs while forming its judgments. As a result, even though each score is independently assigned, the scoring process reflects relative differences in perceived exaggeration between the two pairs.

The assigned scores are discrete values (0, 1, 2, or 3). At first, this may look like many examples will have the same score, and this could make it difficult to rank thousands of samples. In our method, however, the important point is not the absolute value itself but the relative order when two summaries are scored together. Even if both summaries get the same score, the final ranking is decided by combining many pairwise results in the sorting and ensemble process. Sometimes cyclic cases can appear (e.g., A > B, B > C, and C > A), but these conflicts are reduced by the ensemble sorting, which averages results over multiple runs. Because of this design, using discrete scores does not cause problems in ranking and still gives a reliable basis for making continuous exaggeration scores.

An example of the prompt is shown in Appendix A (Figure 5).

#### 4.1.3 Sorting and Scoring Procedure

We used a MergeSort-based sorting algorithm, where the model's output scores were used to compare and sort samples. To make the results more

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

reliable, we repeated the sorting process four times, each time with a different random order of the 1,000 samples.

To calculate the final exaggeration score for each sample, we did the following:

1. Averaged the rank positions from the four runs
2. Normalized the average rank to a value between $[0, 1]$

This process gave each sample a single exaggeration score, which can be used for further tasks such as supervised learning or interpretation analysis.

#### 4.1.4 Evaluation Focus

We evaluated the dataset of exaggeration scores from two main perspectives:

**Stability** Do the scores remain consistent across different sorting runs?

**Validity** Do the scores reflect clear and understandable differences between exaggerated and non-exaggerated summaries?

### 4.2 Stability of Score Assignment

In this experiment, we aimed to check whether the ranking results obtained through LLM-based pairwise comparisons are stable. If the pairwise judgments change a lot between different runs, or if the MergeSort algorithm is very sensitive to the initial order of inputs, the final exaggeration scores may not be reliable. Although the ensemble method introduced in Section 3.4.2 is designed to improve robustness, its effectiveness still depends on the basic stability of the sorting results.

To evaluate this stability, we ran the sorting process four times using different random shufflings of the 1,000 article-summary pairs. For each pair, we recorded its final rank from each run and calculated the standard deviation of its four rankings.

For example:

- Sample A was ranked 2nd, 3rd, 2nd, and 2nd $\rightarrow$ Standard deviation: 0.43 $\rightarrow$ ✓ Stable
- Sample B was ranked 4th, 6th, 1st, and 2nd $\rightarrow$ Standard deviation: 1.92 $\rightarrow$ ✗ Unstable

A smaller standard deviation means the rankings are more consistent across runs, which indicates higher stability.

On average, the standard deviation across all 1,000 samples was 201.8. This means that each sample's rank changed by about 202 positions on average, out of 1,000. In other words, the variation corresponds to roughly one-fifth of the entire ranking range. This result suggests that the stability of the LLM-based sorting process is moderate. Therefore, the underlying ranking mechanism can be considered reliable enough for generating scores, which supports the use of the ensemble-based scoring approach.

We also analyzed the relationship between the final exaggeration score and the ranking variance. Figure 3 shows the final exaggeration score on the x-axis and the standard deviation from the four runs on the y-axis. The orange horizontal line in the figure shows the mean standard deviation ($\bar{\sigma} = 201.8$). Samples below this line are more stable than average, and samples above this line are less stable. According to the figure, samples with scores near 0 or 1 (meaning the least or most exaggerated summaries) tended to have smaller standard deviations, showing that the model's judgments were more consistent in those cases. On the other hand, samples that were ranked around the middle showed larger variations, which suggests that it was harder for the model to assess their level of exaggeration clearly.

One possible way to reduce the impact of this middle-range instability is to use binning instead of relying only on fine-grained continuous scores. For example, dividing the range into broad categories such as "low," "medium," and "high" exaggeration could provide a more robust basis for downstream applications.

These findings suggest that LLM-based ranking is stable and reliable for identifying summaries that are clearly exaggerated or not exaggerated. However, the low variance near the edges of the score range may also come from boundary effects, because items close to 0 cannot be rated much lower and items close to 1 cannot be rated much higher. In future work, it will be important to study the stability of scores in the middle range, where the level of exaggeration is less clear, to understand the limits of our method better.

Based on these findings, we conclude that LLM-based ranking is especially stable and reliable for identifying extremely exaggerated or non-exaggerated summaries. However, rankings for moderately exaggerated summaries may involve more uncertainty.

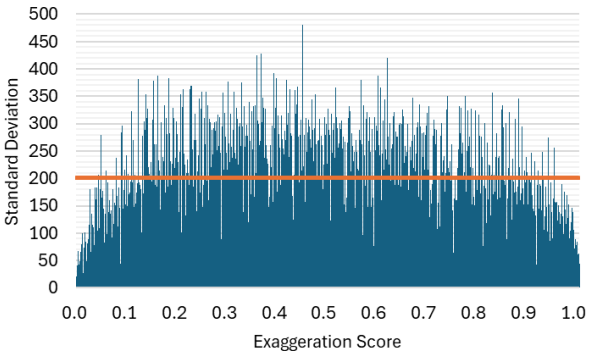


Figure 3: Standard deviation of ranks across four sorting runs, plotted against the final exaggeration score.

### 4.3 Validity of Score Assignment

In Section 4.2, we showed that the ranking results based on LLM comparisons are stable, especially when the article-summary pairs are clearly exaggerated or not exaggerated. Based on this, we now look into whether the exaggeration scores are meaningful and consistent with what people usually expect.

To do this, we looked at examples from both ends of the score range: summaries with scores close to 0.0 (least exaggerated) and those close to 1.0 (most exaggerated).

For instance, in a typical case with a score of 0.0, the article was about pre-season ranking results of a university sports team. The summary gave a short and accurate explanation of the rankings. It kept a neutral tone and didn't use any emotional or judgmental language. It covered the original article's content properly without leaving out or exaggerating any parts (see Figure 7 in Appendix B).

On the other hand, a summary with a score of 1.0 came from an article about economic and political instability across several countries in Eastern Europe. However, the summary focused only on the relatively optimistic outlook of one specific region, which was not the main point of the article. It also used strong phrases like "shielded from disaster" or "the full might of the Swedish state" which were not mentioned in the article. Because of this, the summary gave a misleading feeling of safety, which was very different from the article's more careful and balanced tone (see Figure 8 in Appendix B).

These findings suggest that summaries with low exaggeration scores usually keep the original content's meaning and neutrality, while high-scoring summaries often stress certain opinions, use emotional words, or focus on narrow parts of the article in a way that changes the original message.

So, we conclude that the exaggeration scores are not only statistically reliable but also meaningful in terms of content. The highest and lowest score cases match what human readers would consider most or least exaggerated, which shows that our score is useful for understanding how much a summary exaggerates.

### 4.4 Validating the Exaggeration Score Using Artificial Summaries

To check whether our exaggeration score properly reflects the level of exaggeration, we conducted an experiment using artificially created exaggerated summaries. In this experiment, we examined whether our score matches binary exaggeration labels: exaggerated and non-exaggerated.

To build this labeled dataset, we partially followed the method used in JExnoS (Iwamoto and Shimada, 2024), especially the part where exaggerated summaries were generated using GPT-4o. Based on summaries from the CNN/DailyMail dataset (Nallapati et al., 2016), GPT-4o was instructed to rewrite each summary by adding one exaggerated expression, without introducing any factual mistakes. The prompt we used for this task is shown in Appendix A (Figure 6).

We then calculated exaggeration scores for both the original and exaggerated summaries using our proposed method. As we expected, the exaggerated summaries received higher scores than the originals. Specifically, in a set of 1,000 samples, the average score for non-exaggerated (original) summaries was 0.42, while exaggerated summaries had an average of 0.58. This result indicates that our score can clearly distinguish between the two types.

To help show the difference more clearly, we made a visualization of how the two classes are distributed based on the exaggeration scores, as shown in Figure 4.

The figure has two parts. The top part is a horizontal line that shows all 1,000 summary samples. Each small vertical line represents one summary. Red lines show exaggerated summaries, and blue lines show non-exaggerated ones (original summaries). From left to right, the exaggeration score increases from 0.0 to 1.0. We can see that the red lines are mostly on the right, and the blue ones are mainly on the left.

The bottom part is a histogram that shows how many summaries are in each score range. The red bars show exaggerated summaries, and the blue

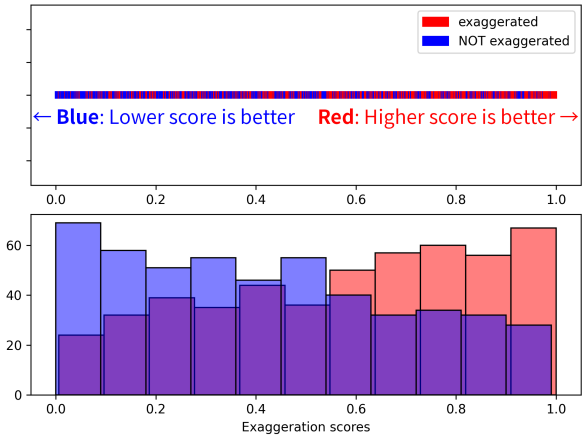


Figure 4: Distribution of Exaggeration Scores for Labeled Summaries.

bars show non-exaggerated ones. As the score goes up, the red bars get higher, and the blue bars get lower. This clear trend means that our exaggeration score matches well with the binary labels, with higher scores usually indicating more exaggerated content.

In summary, this figure gives strong evidence that our exaggeration score reflects the actual degree of exaggeration. It can smoothly separate exaggerated and non-exaggerated summaries across a continuous score range in an easy-to-understand way.

These findings suggest that our exaggeration score not only reflects intuitive exaggeration levels but also matches external exaggeration labels. This supports the reliability of the score and its potential usefulness in tasks such as bias detection and automatic summary checking.

## 5 Conclusion

In this study, we proposed a new method for measuring the degree of exaggeration in news article summaries. Our approach uses relative pairwise comparisons performed by an LLM to create a dataset of ranked summary pairs. Then, we applied an ensemble-based MergeSort algorithm to assign continuous exaggeration scores to each summary.

We conducted several experiments to evaluate our method from three main perspectives:

**Stability** We showed that the rankings produced by the LLM-based comparison function are consistent across multiple runs. This suggests that the exaggeration scores are stable and not strongly affected by random factors.

**Validity** We found that summaries with very high or very low exaggeration scores have linguistic features that match human intuition. These features include emotional language, strong assertions, and distorted information.

**Sensitivity** We confirmed that our score reacts properly to artificial exaggeration. When we intentionally rewrote summaries to add exaggeration, the scores increased, correctly reflecting the added bias.

These results indicate that our exaggeration score is a reliable and interpretable measure for analyzing exaggeration in news summaries. By turning subjective impressions into a continuous, data-driven metric, our method allows for more detailed and scalable analysis of media bias, sensationalism, and misinformation.

To further advance this research, there are several directions for future work:

**Score calibration** Our current method generates relative scores. However, matching these to absolute human judgments or real-world consequences remains a challenge.

**Practical applications** We plan to explore the use of the exaggeration score in real tasks, such as training models to detect exaggeration, assessing the credibility of headlines, and improving the quality of automatic summarization.

We hope that our work helps deepen the understanding of subtle framing effects in summaries and contributes to more responsible communication in today's digital information environment.

## Limitations

Although our proposed method shows encouraging results in evaluating exaggeration in news summaries, there are still several limitations that need to be addressed in future research.

### Non-linear perception of exaggeration

In our current method, we assume a linear relationship between the least and most exaggerated summaries when assigning scores. However, how people actually perceive exaggeration may not follow a simple linear pattern. For example, certain types of exaggeration, especially those that use emotionally strong or misleading language, might affect readers' impressions more strongly than others. Therefore, it may be useful to apply non-linear

scaling techniques, such as logarithmic or sigmoid functions, which are informed by user studies or cognitive models. This could help make the scoring system more consistent with human perception.

### Human evaluation

Another limitation of this study is the absence of systematic human evaluation. While our LLM-based comparisons provide stable and interpretable exaggeration scores, the reliability of these scores should be further validated by human judgments. However, it is often difficult for human annotators to assign absolute exaggeration scores, such as deciding whether a summary deserves a score of 0.7 or 0.8. A more practical approach is relative evaluation, where annotators compare two summaries and decide which one is more exaggerated. Such pairwise judgments are easier for humans and can later be aggregated into consistent rankings. In future work, we plan to design small-scale human studies based on this relative evaluation framework to strengthen the validity of our dataset.

### Cultural and linguistic bias

In this study, we used datasets such as Newsroom and CNN/DailyMail, which mainly include English news articles from U.S.-based media. Because of this, the ways exaggeration appears in these summaries may reflect cultural characteristics or journalistic styles that are specific to the United States. As a result, it may be difficult to apply the same method to other languages, cultures, or media environments. To improve the general usefulness of the method, it would be important to apply it to multilingual or culturally diverse datasets in future work.

## References


William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Keisuke Iwamoto and Kazutaka Shimada. 2024. Dataset construction and verification for detecting factual inconsistency in japanese summarization. In *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 243–248.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.


## A Prompts Used in This Study

In this section, we show the prompts used in our experiments. The first prompt is for pairwise comparison of summaries to judge which is more exaggerated. The second prompt is for generating new exaggerated summaries from original ones.

```
Please act as a fair judge and evaluate
the level of exaggeration in the
summaries of two news article pairs (A
and B). Each pair includes a full
article and its summary. Your task is to
 score how much the summary exaggerates
the content of the original article.
Use the following 4-point scale:
- 0: No exaggeration
- 1: Slight exaggeration
- 2: Moderate exaggeration
- 3: Strong exaggeration

When scoring, consider the following:
- Exaggerated numbers or statistics
- Making the subject or event seem more
important than it is in the original
article
- Use of meiosis (understatements) or
litotes (negation-based softening)
- Stylistic choices that may change the
overall tone or impression of the
article

First, explain your reasoning for each
score, using specific examples from the
summaries.
Be as objective as possible. Do not be
influenced by the length or order of the
 summaries.
```

Figure 5: Example of the prompt used for pairwise comparison in the exaggeration scoring task.

```
You are given a news article and its
summary.
Rewrite the summary to generate an
exaggerated version by introducing only
one overexpression.
This overexpression should not cause
clear factual inconsistencies, such as
contradictions or incorrect numbers.

After generating the exaggerated summary
, provide the following:
- Overexpressed Summary: The rewritten
exaggerated version of the summary.
```

Figure 6: Example of the prompt used for generating exaggerated summaries.

## B Summary Examples for Extreme Scores

This section shows two summary examples from our dataset. One has an exaggeration score of 0.0 (no exaggeration), and the other has a score of 1.0 (strong exaggeration). These examples help readers understand what kind of summaries get extreme scores.

Northern Arizona men's basketball is picked to finish third in the Big Sky Conference by conference coaches and second in a preseason media poll.

Figure 7: Summary example (score 0.0). Article: Fox Sports, Oct 23, 2014.[2]

The Baltic trio of Latvia, Lithuania, and Estonia are lucky. At the end of the day, they can count on Swedish banks and the full might of the Swedish state to shield them from economic disaster.

Figure 8: Summary example (score 1.0). Article: The Telegraph, Jul 6, 2009.[3]

[2] https://web.archive.org/web/20150403043731/http://www.foxsports.com/arizona/story/northern-arizona-picked-in-top-3-in-big-sky-basketball-102314

[3] https://web.archive.org/web/20220817132949/https://www.telegraph.co.uk/finance/newsbysector/banksandfinance/5760816/Bulgarian-stress-test-for-the-Balkans.html