References Matter: Investigating the Impact of Reference Set Variation on Summarization Evaluation

Silvia Casola[*] **Yang Janet Liu**[*] **Siyao Peng**[*]
Oliver Kraus **Albert Gatt** **Barbara Plank**
MaiNLP, LMU Munich   Munich Center for Machine Learning (MCML), Germany
Department of Linguistics, University of Pittsburgh, USA
NLP Group, Department of Information and Computing Sciences, Utrecht University, Netherlands
{s.casola, siyao.peng, o.kraus2, b.plank}@lmu.de,
jal787@pitt.edu, a.gatt@uu.nl

Abstract

Human language production exhibits remarkable richness and variation, reflecting diverse communication styles and intents. However, this variation is often overlooked in summarization evaluation. While having multiple reference summaries is known to improve correlation with human judgments, the impact of the reference set on reference-based metrics has not been systematically investigated. This work examines the sensitivity of widely used reference-based metrics in relation to the choice of reference sets, analyzing three diverse multi-reference summarization datasets: SummEval, GUMSum, and DUC2004. We demonstrate that many popular metrics exhibit significant instability. This instability is particularly concerning for n-gram-based metrics like ROUGE, where model rankings vary depending on the reference sets, undermining the reliability of model comparisons. We also collect human judgments on LLM outputs for genre-diverse data and examine their correlation with metrics to supplement existing findings beyond newswire summaries, finding weak-to-no correlation. Taken together, we recommend incorporating reference set variation into summarization evaluation to enhance consistency alongside correlation with human judgments, especially when evaluating LLMs.

1 Introduction

Human-written texts vary widely in terms of length, style, communicative intent, lexical/syntactical choices, and numerous other dimensions (Giulianelli et al., 2023; Liu and Zeldes, 2023; Rezapour et al., 2022; Baan et al., 2023). Such variation poses a significant challenge in the evaluation of summarization systems (Lloret et al., 2018; Celikyilmaz et al., 2021). Traditional summarization metrics typically rely on comparing

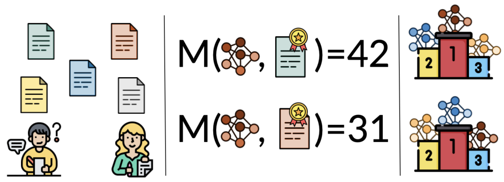


Figure 1: **Human-written summaries are diverse.** Using a human-written reference over another makes evaluation metrics fluctuate, and affects model ranking.

system outputs to one or more references, treating these references as a "gold standard". Although the limitations of reference-based metrics have long been acknowledged (Rankel et al., 2013; Louis and Nenkova, 2013; Reiter, 2018; Peyrard, 2019; Fabbri et al., 2021; Goyal et al., 2023), they remain widely popular due to their simplicity, low compute requirements, relative ease of adaptation to different languages, and reproducibility.

The assumption behind the use of reference-based metrics is that system outputs that are more similar to the reference(s) are better, due to their "human-likeness" (Gehrmann et al., 2023). However, the significant variation in human-written summaries implies that evaluating system outputs against a single or limited set of references has inherent drawbacks. Previous research has extensively looked at correlations between metrics and human judgments in summarization (Forde et al., 2024; Mondshine et al., 2025), further exploring the use of multiple references to improve such correlations (Lin, 2004; Belz and Reiter, 2006; Fabbri et al., 2021; Tang et al., 2024) as well as the interpretability and efficiency aspects of such automatic metrics (Liu et al., 2023b). However, a much less studied question is **the extent to which automatic metrics are sensitive to the *choice* of human-written reference summaries**, as shown in Figure 1. In other words, are these metrics stable across different plausible gold-standard refer-

[*] Equal contribution.

Proceedings of the 18th International Natural Language Generation Conference, pages 274–291
October 29–November 2, 2025.

ences? If metric scores vary significantly with the selected reference(s), this variation calls into question the reliability of many evaluation practices in the field.

In this work, we quantify the impact of reference choice on automatic evaluation metrics for summarization. Our contributions are as follows:

[1] **We investigate how different reference sets affect system rankings**. We show that system rankings based on n-gram-matching metrics (e.g., ROUGE) strongly depend on the choice of the reference(s), undermining the reliability of model comparisons. However, rankings based on more semantically-oriented metrics exhibit greater stability.

[2] **We examine the robustness of widely-used reference-based metrics for summarization at the instance and dataset level.** Our analysis reveals that the variation in scores introduced by the choice of reference on a dataset often exceeds the variation observed in state-of-the-art (SOTA) models.

[3] **We collect new human judgment scores** on Large Language Model (LLM) outputs for the genre-diverse GUMSum (Liu and Zeldes, 2023) dataset. We use these data **to reassess the correlation between automatic metrics and human judgments**, complementing earlier SummEval evaluations (Fabbri et al., 2021), which were limited to pre-LLM models and newswire data. We find that correlations tend to increase with the number of references, and that the metric with the highest correlation varies depending on the evaluation dimension *and* the number of references.

Our analysis reveals that few metrics tend to show both reasonable correlation with human judgments *and* robustness to the reference sets, especially when scoring LLM outputs.

The code is available at `https://github.com/mainlp/references-matter`.

2 Related Work

Summarization Evaluation. Recent advances in Natural Language Generation (NLG) have significantly enhanced the development of summarization systems. However, their evaluation remains an open problem (Celikyilmaz et al., 2021; Goyal et al., 2023). Summarization evaluation metrics are broadly categorized into reference-based and reference-free (Lloret et al., 2018). Reference-based metrics compare system outputs to human-written reference summaries, relying on methods such as n-gram overlap (Lin, 2004; Papineni et al., 2002), embedding similarity (Ng and Abrecht, 2015; Zhao et al., 2019; Zhang et al., 2020), or model-based techniques (Peyrard et al., 2017; Scialom et al., 2019; Yuan et al., 2021). In contrast, reference-free summarization metrics do not assume a gold standard (Yuan et al., 2021; Vasilyev et al., 2020; Gao et al., 2020; Gigant et al., 2024). More recently, growing research leverages LLMs as evaluators, with or without references (Song et al., 2024; Li et al., 2024). In the LLM-as-judge paradigm, evaluations are typically based either on prompting the model to provide judgments (Liu et al., 2023a; Bavaresco et al., 2025), or on using its generative probabilities directly (Fu et al., 2024).

Metrics Meta-Evaluation. Meta-evaluation of summarization metrics typically focuses on the extent to which they can be used as a proxy for human evaluation. Reiter and Belz (2009) examined the validity of automatic scores for NLG tasks, while Rankel et al. (2013) focused on ROUGE and its correlation with humans. Peyrard (2019) showed that metrics with reasonable correlation on lower-quality outputs tend to diverge when output quality increases. Caglayan et al. (2020) demonstrated the idiosyncrasies of automatic evaluation metrics, noting that high correlation with human judgments is not sufficient to characterize their reliability. Fabbri et al. (2021) performed a large-scale meta-evaluation of summarization metrics, and found that most metrics have low correlation with human judgments on *coherence*, while *relevance* is weakly or moderately correlated. Mondshine et al. (2025) performed a meta-analysis of reference-based, reference-free, and LLM-based summarization metrics focusing on eight languages from four typological families, showing low correlation. They also observed that off-the-shelf LLMs as judges still lag behind other metrics.

While most existing research focused on correlation to human scores, Tang et al. (2024) addressed the challenge of evaluation when a limited number of references is available. They proposed leveraging LLMs to diversify the references, expanding the evaluation coverage and improving the correlation with humans. Their results show that increasing the number of references significantly enhances the reliability of existing evaluation metrics in terms of correlation. However, since LLM outputs tend to show less variability and follow distinct patterns compared to human-produced content (Giulianelli et al., 2023; Guo et al., 2024; Shur-Ofry et al., 2024; Reinhart et al., 2025), relying on them to replace human references might introduce biases. Beyond summarization, evaluation of reference variability has also been conducted in tasks such as machine translation (Castilho, 2020; Popović, 2021; Wu et al., 2025) and image captioning (Yi et al., 2020).

3 Experimental Setup

To quantify the impact of human-written references on the scores of automatic metrics, we leverage multiple elements. For datasets, we use SummEval (Fabbri et al., 2021), GUMSum (Liu and Zeldes, 2023), and DUC2004 (Dang and Croft, 2004), which contain multiple human-written summaries (§3.1), to assess how different reference summaries affect metric performances. Next, to assess summarization models, we use the existing outputs provided by Fabbri et al. (2021) for SummEval. As these outputs predate LLMs, we additionally collect outputs using LLMs (§3.2) for all three datasets. Lastly, to compute the correlations with humans, we use the human judgments available in SummEval and gather new human ratings for GUMSum on both human and LLM-generated summaries (§3.3). We prioritize GUMSum over DUC2004, as it includes multiple genres beyond news data. Our metric selection is outlined in §3.4. Details on data licensing and codebase are provided in Appendix A.

3.1 Human-written Summaries

SummEval (Fabbri et al., 2021) is built on top of CNN/DM (Hermann et al., 2015; Nallapati et al., 2016), containing news articles and human-written highlights. The original authors selected 100 instances from the test set and supplemented the existing highlights with ten additional reference summaries per instance, obtained via crowdsourcing (Kryscinski et al., 2019).

GUMSum (Liu and Zeldes, 2023) contains summaries created following general and genre-specific guidelines[1] to function as a substitute for the source (Nenkova and McKeown, 2011). We focus on the 48 documents in the dev and test sets, which contain five human-written summaries each (Lin and Zeldes, 2025) across 12 genres.

DUC2004 Task1 (Dang and Croft, 2004) consists of 489 news documents, most with four references. The guidelines allow the summaries to be in the form of short sentences or lists of keywords.[2] DUC2004 references are thus extremely concise (only up to 75 characters). The dataset has played a significant role in summarization research, being part of the annual TREC conference evaluation.

Table 1 provides an overview of the three datasets. We treat all human references in the three datasets as "gold" since they were either authored by experts or validated through review.

3.2 Model Outputs

Fabbri et al. (2021) collected model outputs for SummEval from 24 extractive and abstractive summarization systems, which were SOTA between 2017 and 2019. We focus on the 16 models for which they provided human judgments.

For all datasets, we also include summaries generated by contemporary LLMs. This is crucial given that prior studies demonstrated that evaluation metrics often show lower correlation with high-quality outputs (Peyrard, 2019; Alva-Manchego et al., 2021). Below, we report a similar pattern for LLMs (§4.4). For consistency purposes, we follow Lin and Zeldes (2025) and use Llama3-3B-Instruct (Hermann et al., 2015), Qwen-2.5-7B-Instruct (Yang et al., 2025), Claude-3.5 (Anthropic, 2024), and GPT-4o (OpenAI, 2024). For each LLM, we generate a single summary. We emphasize LLM variety over multiple generations. Details on the generation parameters and prompts are reported in Appendix A.

3.3 Human Judgments

SummEval (Fabbri et al., 2021) contains expert judgments that assess summaries based on four criteria: *coherence*, *consistency*, *fluency*, and *relevance*, using a Likert scale of 1-5 (Likert, 1932).

To measure how well automatic metrics align with human judgments beyond the news domain, and to study whether findings on pre-LLM models align with those on LLM outputs, we collect a new set of human judgments using the same criteria on the 48 GUMSum documents. We hired

[1] https://wiki.gucorpling.org/gum/summarization

[2] https://duc.nist.gov/duc2004/tasks.html

dataset	#doc	genre	references			outputs	
			#sums	#chars	#toks	model outputs	human judgments
SummEval	100	news	1+10	226.3	43.1	Fabbri et al. (2021) + 4 LLMs	Fabbri et al. (2021)
GUMSum	48	12 genres	5	291.3	52.1	4 LLMs	collected in this work
DUC2004	489	news	4	70.0	11.9	4 LLMs	n/a

Table 1: **Multi-reference summarization datasets.** #sums indicates the number of human-written references per instance. We generate outputs using four LLMs and collect a new set of human judgments for GUMSum.

Summarizer	Coh.↑	Con.↑	Flu.↑	Rel.↑	best↑	worst↓
claude	**4.75$_{0.45}$**	4.48$_{0.65}$	**4.82$_{0.42}$**	**4.26$_{0.71}$**	**0.17$_{0.38}$**	**0.03$_{0.18}$**
gpt4o	4.42$_{0.61}$	**4.61$_{0.63}$**	4.74$_{0.45}$	4.07$_{0.65}$	0.11$_{0.32}$	0.12$_{0.32}$
Qwen	4.66$_{0.56}$	4.33$_{0.84}$	4.68$_{0.52}$	4.23$_{0.75}$	0.16$_{0.37}$	0.12$_{0.33}$
Llama3	4.71$_{0.51}$	4.14$_{0.97}$	4.78$_{0.42}$	4.09$_{0.86}$	0.12$_{0.32}$	0.20$_{0.40}$
humans	4.54$_{0.57}$	4.48$_{0.71}$	4.70$_{0.51}$	4.22$_{0.69}$	0.09$_{0.28}$	0.10$_{0.31}$

Table 2: **Human judgments on GUMSum: LLM vs. human-written summaries.** Above-human performances are highlighted in blue.

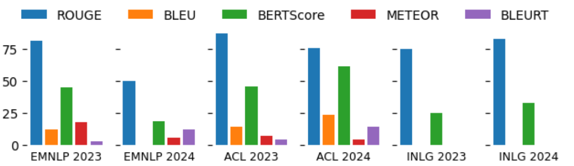


Figure 2: Percentage of papers about summarization that use common reference-based metrics.

three Master's students in Computational Linguistics and tasked them to evaluate four LLM outputs (§3.2) and five human references (§3.1), following Fabbri et al. (2021)'s criteria. LLM-generated and human-written summaries were anonymized and shuffled. We also asked the evaluators to pick one best and one worst summary for each document.

Table 2 reports the results. Claude scored the best overall. GPT-4o gets the highest *consistency* but the lowest *coherence* and *relevance*, and is thus the least picked LLM. Interestingly, LLM outputs typically receive higher scores than human-written references. In line with previous work, e.g., Zhang et al. (2024), this finding has significant implications for reference-based evaluation and calls into question the use of potentially lower-quality references for assessing high-quality outputs (Noh et al., 2024).

3.4 Evaluation Metrics

We consider several reference-based metrics, chosen to balance popularity and diversity. All metrics range in 0–100. Appendix B provides details.

ROUGE (Lin, 2004) is the most popular summarization metric. ROUGE-N computes n-gram overlap between a hypothesis and the references. ROUGE-L leverages the longest common subsequence, accounting for the word order. With multiple references, ROUGE considers the maximum or the mean of the n-gram overlap (ROUGE$_{max}$ and ROUGE$_{avg}$). We report the F1-score.

BLEU (Papineni et al., 2002) is an n-gram metric primarily used for translation. It is precision-based and incorporates a brevity penalty. With multiple references, the n-gram count is clipped at the maximum count of n-grams in a single reference, and the length of the reference closest in size to the hypothesis is considered.

METEOR (Banerjee and Lavie, 2005) incorporates multiple linguistic aspects, including synonym matching, stemming, and word order, making it more robust in capturing semantic equivalence. While primarily designed for translation, it has also been used to assess summaries. With multiple references, the maximum score is considered.

BERTScore (Zhang et al., 2020) leverages contextual embeddings and considers the cosine similarity between the embeddings of the hypothesis and the reference tokens. With multiple references, the final score is the maximum among the individual scores. We report the F1 score.

BLEURT (Sellam et al., 2020) is a model-based metric that leverages BERT fine-tuned on human judgments. The metric is not designed to handle multiple references; we compute scores for each reference and consider the maximum.

To gain insights into recent metric usage, Figure 2 summarizes the percentage of summarization papers from recent ACL, EMNLP, and INLG proceedings (detailed in Appendix C). We found that reference-based metrics are still the most popular, with ROUGE in 79% of papers, followed by BERTScore (44%). Our preliminary search also shows that LLM-as-a-judge evaluators are mainly (i.e., in 66% of the cases) used without references, placing them outside the scope of this paper.

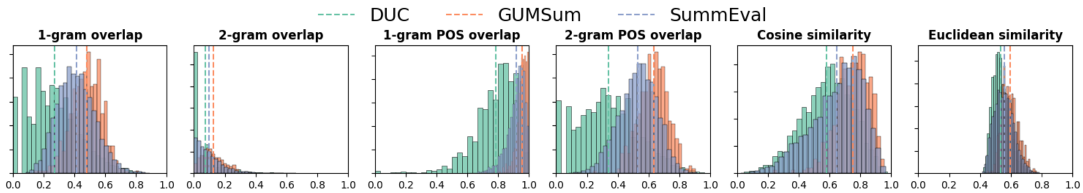


Figure 3: Variation in human-written summaries across datasets, measures inspired by Giulianelli et al. (2023).

4 Reference Variability and Metric Robustness

Reference-based metrics assume that more human-like outputs deserve higher scores. However, human summaries are very diverse. This section examines how metrics fluctuate with different human references. By analyzing metric robustness, we aim to understand how conclusions about models, drawn from reference-based metrics, might change when different sets of human-written references are used, thereby undermining evaluation reliability.

4.1 Human-written Summaries are Diverse

Human-written summaries show substantial diversity. We assess the variability in the multi-reference datasets following Giulianelli et al. (2023). For each pair of human-written summaries for the same instance, we report the lexical similarity (the overlapping distinct n-grams between two strings), the syntactic similarity (the overlap of part-of-speech tag n-grams), and the semantic similarity (the cosine and euclidean similarity between the embeddings of the two strings).

Figure 3 shows these variations. At the dataset level, DUC and SummEval show the lowest similarity among human-written summaries across all dimensions. For GUMSum, summaries are more similar to each other. We hypothesize that this is likely due to the constrained annotation guidelines. It is also worth noting that the similarities revealed here are between different human-written summaries for a given instance as opposed to summaries across genres, for which we still expect significant variations, as demonstrated by Liu and Zeldes (2023). Overall, summaries tend to be similar at the syntactic level, less so at the semantic and lexical level. We also observe that LLM outputs show lower diversity (Appendix D), consistently with previous work (Giulianelli et al., 2023).

4.2 Automatic Metrics Fluctuate Substantially at the Instance Level

Given the diversity in human-written summaries, we quantify metric fluctuation at the instance level when using a different set of human-written references. For a metric M and a set of human-written references $R = \{r_1, r_2, \ldots, r_N\}$, we compute $M(r_i, R - \{r_i\})$. Thus, for each document, we score each human-written summary using all the others as the reference set. Figure 5 exemplifies the observed instance-level variability measured by ROUGE-L_{avg} on the three datasets. For SummEval, we also mark the original reference (scraped highlights, §3.1) in the CNN/DM dataset with a cross. The quality of these scraped references versus the ten later crowd-sourced ones is discussed further in Appendix E.

Scores assigned to human-written summaries are often low. For example, the averaged ROUGE-L_{avg} scores are $28.52_{\pm 5}$, $27.46_{\pm 3}$, $24.88_{\pm 5.3}$ for SummEval, GUMSum, and DUC2004. Given the assumption that human reference summaries are of high quality, metrics should produce high scores. Instead, they do not typically reflect this property.

Human-written reference scores vary widely. Figure 4 summarizes the instance-level variability of the individual scores (in Figure 5) for all evaluation metrics on SummEval (corresponding figures for GUMSum and DUC are in Appendix F). For each metric, we compute the min-max range when scoring human-written references against all the others ($M(r_i, R - \{r_i\})$). Figure 4 shows the histogram of such ranges. **The ranges of variation observed within human-written references are, on average, very high**.

Understanding the magnitude of such a range might not be obvious. For instance, an increase of 10 points of BERTScore (typically scoring in the high range of the scale) might indicate a much larger improvement in performance than an in-

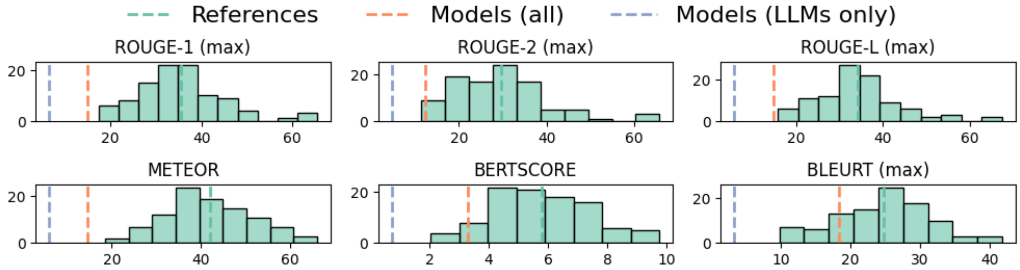


Figure 4: **Ranges of variability at the instance level on SummEval.** For each instance, we compute the range of the scores of the references against the remaining ones. The trends for $\text{ROUGE}_{\text{max}}$ and $\text{ROUGE}_{\text{avg}}$ are similar.

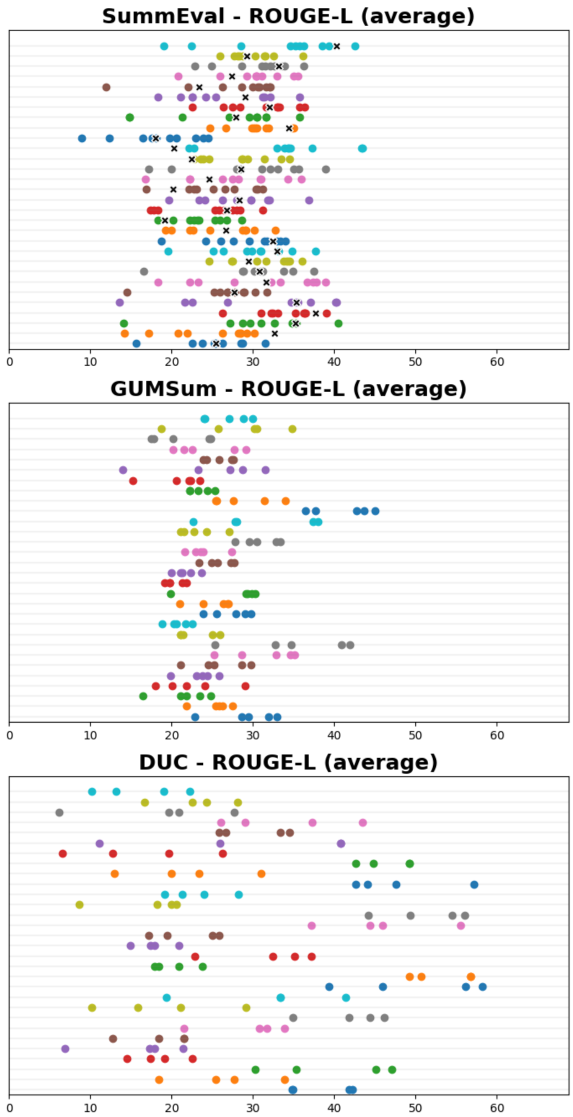


Figure 5: **Instance-level variation for $\text{ROUGE-L}_{\text{avg}}$.** For every document (shown first 30, one per line), we plot the score for every human-written reference against all other references (using the same color per source to aid interpretation). The original CNN/DM reference in SummEval is marked by a cross.

crease of 10 points of ROUGE-1.[3] To contextualize the magnitude of variation for each metric, we also report the performance range of summarization systems. Thus, for a model S, given its output o_i for instance i, we score it through $M(o_i, R)$. Although these values are not directly comparable and should be interpreted with caution due to the use of different reference sets, they help contextualize the magnitude of the results and its potential impact on evaluation. For example, ROUGE-1_{max} assigned to human-written references varies by about 35 points on average (the green dashed line in Figure 4), while the mean range is less than 20 points across all model outputs (orange line), and much lower for LLM outputs (blue line). **Similarly, LLM summaries exhibit much less variability than human references on all metrics.** These findings highlight the significance of variability and suggest that the ranking of summarization models is highly sensitive to the reference set.

4.3 System Ranking Depends on the Reference(s) for N-Gram-Based Metrics

While we observed variability at the instance level, summarization metrics are typically designed to evaluate models across datasets, rather than individual instances. In this section, we investigate to what degree standard summarization metrics can handle the variability observed in human-written references when ranking summarization systems.

Procedure. We sample k human-written references ($k \in [1, N]$, where N is the number of references for each document) from all available references for each instance. We then score the outputs of each summarization system using the same set of references. Given M systems $S_a, S_b, \ldots, S_M$, the metric induces a ranking $S_a \succ S_b \succ \cdots \succ$

[3]The relative dynamics between metrics have been studied by Kocmi et al. (2024) in machine translation.

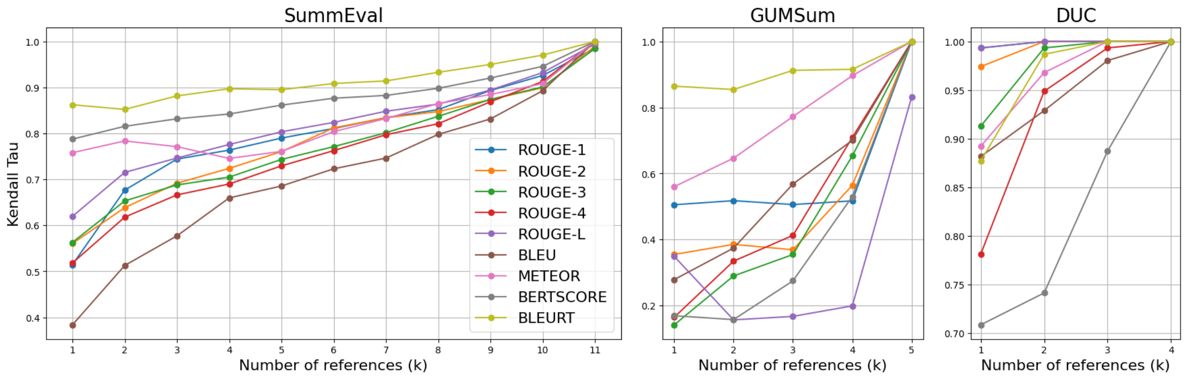


Figure 6: **Rank stability when increasing the number of references**. $\text{ROUGE}_{\text{max}}$ is presented. Note that we use different ranges for the y axis for each dataset to improve readability.

S_M. This process is repeated 100 times, yielding 100 rankings. We compute the pairwise Kendall rank correlation coefficient (Kendall, 1938) between such ranks. High correlation indicates that models are similarly ordered, even when different sets of references are used.[4] Figure 6 reports the average correlation for pairs of ranks for each dataset and metric, from using k human-written summaries as references. $\text{ROUGE}_{\text{max}}$ is shown in Figure 6, and $\text{ROUGE}_{\text{avg}}$ is reported in Figure 12 in Appendix G.

Single Reference. Evaluating with a single reference is common in summarization, as most datasets provide only one human-written summary. Figure 6 (looking at $k = 1$) shows the stability of different metrics with a single reference across the three datasets. We find that **BLEU and ROUGE have very weak to moderate correlation between ranks across different references**. In other words, using two different sets of plausible references would likely lead to different conclusions on relative model performance. We also notice a large variability among the individual pairs of rankings, with some showing negative correlation (refer to Table 4 in Appendix G for results on individual metrics and datasets).

In contrast, **more semantically-oriented metrics show greater stability**. For SummEval, BLEURT shows the highest correlation between ranks, followed by METEOR and BERTScore. BLEURT and METEOR confirm their stability on GUMSum when ranking the LLM outputs. Other metrics (including BERTScore) show low or no correlation on GUMSum, with the exception of ROUGE-1. In all cases, metrics show much higher stability on DUC, for which all average correlations are above 0.7. We speculate that high stability might stem from an artifact introduced by the short summary length required by the guidelines.

In summary, n-gram-matching metrics, though simple, are highly reference-dependent, undermining consistent model evaluation, while semantically-oriented ones show greater stability. Therefore, **we recommend always using model-based metrics in benchmarks with a single reference.** When cost is a factor, METEOR might offer a good balance of stability and affordability.

Multiple References. When scoring model outputs against a set of $k > 1$ randomly sampled references, we observe that **the correlation between rankings obtained with different human-written references generally improves with an increased number of references**. This increased stability is expected and in line with similar findings that associate a larger number of references with a higher correlation with humans (Lin, 2004).

However, the **stability varies by metric**. ROUGE (especially $\text{ROUGE}_{\text{max}}$) and BLEU tend to have low correlation between ranks. As an example, the $\text{ROUGE}_{\text{max}}$ scores require 5-10 references to reach a level of stability that is comparable to that of BERTScore on SummEval with a single reference. $\text{ROUGE}_{\text{avg}}$ has a better stability than $\text{ROUGE}_{\text{max}}$, especially with a larger set of references. For example, on SummEval, $\text{ROUGE-L}_{\text{avg}}$ has higher stability than BERTScore for $k > 3$, while on GUMSum, $\text{ROUGE-2}_{\text{avg}}$ is the second most stable metric for $k > 3$. On all datasets, BLEURT and METEOR remain stable even with a single reference, with METEOR showing stability despite its simplicity.

[4] Note that, for $k > 1$, references in different reference sets might overlap, artificially increasing the observed correlation between rankings.

In general, trends on SummEval are clearer and simpler to interpret than the other two datasets. We speculate that this is due to the larger number of models used (16 pre-LLM models+4 LLMs on SummEval vs 4 LLMs on GUMSum and DUC). BLEURT, METEOR, and BERTScore show the highest stability, while n-gram-based metrics show low to average correlation between ranks even when multiple references are used. The cases of GUMSum and DUC2004 are more complex to interpret and might be less meaningful given fewer model outputs (i.e., only four LLM outputs, which might increase the observed noise). For GUMSum, BLEURT continues to show high inter-rank correlation, with METEOR being the second most stable. BERTScore, on the other hand, shows poor stability. Similar to the case with $k = 1$, on DUC2004, all metrics show high stability, likely due to summaries being very short, as dictated by the guidelines.

4.4 Correlation with Human Judgments

In addition to stability, automatic metrics should correlate with humans. We compute correlations for SummEval and GUMSum, for which we have human judgments,[5] at the instance and system level as the number of references k increases.[6]

Instance-level Correlation. Figure 7 reports the instance-level correlation for SummEval (top) and GUMSum (bottom), respectively, versus the number of references. We show ROUGE$_{\text{max}}$; corresponding figures using ROUGE$_{\text{avg}}$ are in Appendix H.

We notice **weak-to-no correlation** on both datasets. All correlations are generally higher on SummEval (where we consider outputs from the pre-LLM era) than on GUMSum (where we consider LLMs), in accordance to previous work showing that **correlation with human judgments decreases as the quality of the outputs improves** (Peyrard, 2019). Additionally, reference-based evaluation itself could be problematic for very high-quality outputs when references are of worse quality than outputs (see Table 2, where model outputs are often scored on par with, or higher than, references) as also argued by Goyal et al. (2023); metrics might also not be sensitive enough for outputs with more similar quality. The observed low correlation could be motivated by the low IAA in the GUMSum human judgments.

For SummEval, increasing the number of references consistently leads to better correlation. This effect vanishes on GUMSum, where a larger reference set leads to no effect or slightly lower correlation. For SummEval, BERTScore shows the highest correlation on all dimensions but consistency, for which METEOR and ROUGE$_{\text{avg}}$ are better proxies. Notice how the best metric in terms of correlation with human judgment depends on the considered criterion *and* the available number of references: BLEURT, for example, typically has low correlation when considering one reference only, performing worse than ROUGE. However, its performance improves when more references are considered, surpassing the scores of n-gram-based metrics.

System-level Correlation. System-level correlation is generally higher than instance-level correlation on SummEval; however, many criteria still show weak to moderate correlation when one or very few references are included. In most cases, such correlation tends to improve with the number of references. This is not the case for ROUGE$_{\text{max}}$, especially when considering *consistency*. The full results are provided in Figure 14 in the Appendix H. GUMSum is excluded from this analysis due to the small number of systems available.

5 Conclusions

In this work, we have investigated how reference sets impact the reliability of reference-based summarization metrics. Our analysis across three multi-reference datasets reveals that, despite their popularity, token-matching metrics such as ROUGE are highly sensitive to the reference(s). This sensitivity leads to instability in system rankings, particularly when only a small number of references are available, which is typical in summarization datasets. In these situations, we thus recommend avoiding such metrics, echoing earlier calls for caution (Schmidtova et al., 2024), in favor of model-based or reference-free alternatives.

We find that increasing the number of reference summaries consistently improves both the stability of metric scores and their alignment with human judgments. This might be explained by the possibility of representing a larger human diver-

[5] For SummEval, we use the 16 models studied by Fabbri et al. (2021); for GUMSum, the four LLMs.

[6] For example, when considering two references, we consider the sets ($\binom{N}{2}$) of human-written references, where N is the total number of references. We compute the scores using such references as gold standard, and report the mean.

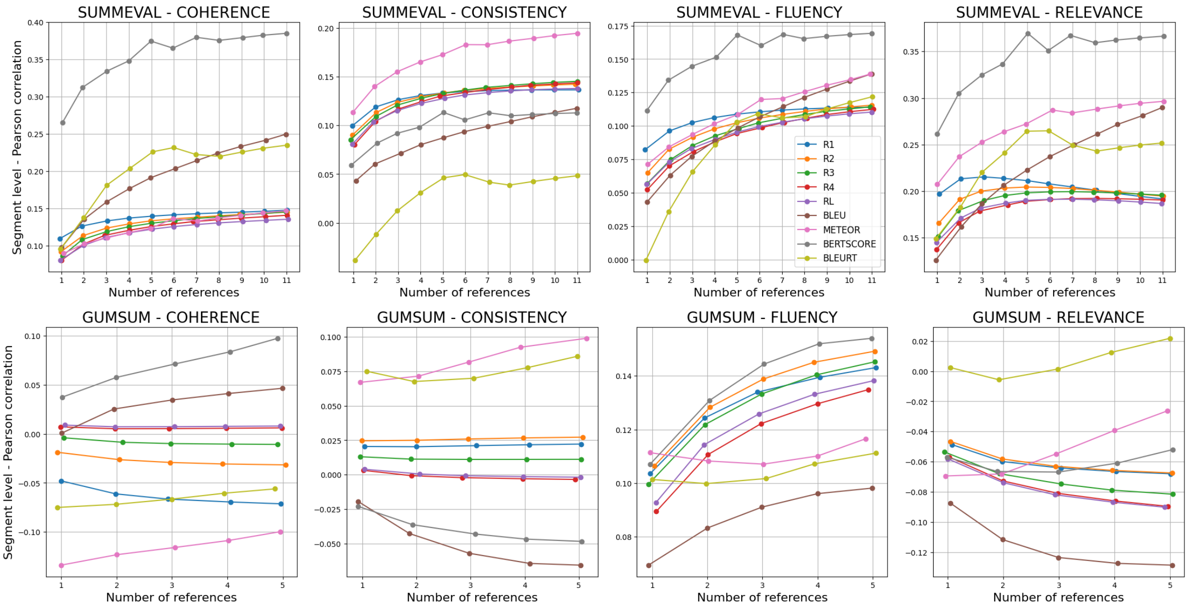


Figure 7: Pearson correlation at the instance level on SummEval (top) and GUMSum (bottom).

sity from the one hand, and from that of limiting annotator bias on the other. In these conditions, n-gram-based metrics such as ROUGE-N become more reliable.

Our findings highlight the need to incorporate reference set variation into evaluation frameworks. Future metric development should explicitly account for this variation. While we have not specifically studied the dimensions of diversity in human references besides their lexical, syntactical, and semantic variation from Giulianelli et al. (2023), we believe this is an important area of investigation. We speculate that the main challenge would be to collect a reference set that is "diverse enough" to represent human production for a fixed number of references. Future work in this direction needs to identify and characterize the relevant dimensions of diversity. These dimensions might be lexical, stylistic, intentional, or even sociolinguistic (cutting across the previously mentioned dimensions, Grieve et al. 2025). How to collect such references and whether specific guidelines should be adopted is also an open problem. Future work should also explore the role of genre.

We thus advocate for the creation of larger, more diverse multi-reference datasets, as well as for metric designs that are inherently robust to reference variability. Such efforts will be key to ensuring fairer and more reliable and human-aligned evaluation practices in summarization in the LLM era and beyond.

Limitations

While our study highlights challenges posed by reference set variation in summarization evaluation, it also comes with several limitations.

Although we focus on multi-reference datasets—SummEval, DUC2004, and GUMSum—such datasets remain relatively small. This reflects a broader limitation in current meta-evaluation practices, where multi-reference resources are the norm despite their limited scale.

Our analysis primarily targets standard evaluation criteria such as *coherence*, *consistency*, *fluency*, and *relevance*. While these are widely adopted and established, they do not capture all the nuances of summary quality, especially when the source texts are genre-diverse. More fine-grained human annotations and task-specific dimensions (e.g., factuality for news, stance for opinion pieces) would allow a deeper understanding of metric behavior under reference variation.

While we assess metric robustness using four systems (all of which are LLMs as opposed to the case for SummEval, where 16 pre-LLM supervised systems are available) on the GUMSum and DUC2004 datasets, the relatively small number of models limits the generalization of the conclusions we can draw about system-level stability. Future work should include a larger and more diverse set of systems to better assess generalizability.

Moreover, using reference-based metrics for LLMs outputs (and generally, very high-quality outputs) has been questioned, especially when considering low-quality (e.g., scraped) references. While we focus on high-quality human-written references especially collected and checked for quality, we find that LLM-outputs are scored higher than references in our human evaluation campaign. We want to point out that the paper does not advocate to use reference-based metrics in such context; rather, it aims at shading lights on the limitation of the current evaluation practices, including the use of references in the cases in which the outputs might have higher quality, and its impact on stability and correlation. We also advocate for more research on the similarities and differences between LLM- and human-written summaries, to understand to which extent the use of LLM output as references could improve evaluation or rather introduce unwanted and largely unknown biases.

Lastly, given our efforts to use contemporary LLMs, there remains a potential risk of data contamination. Since some of these datasets may have been seen during pretraining, this could affect both the outputs generated by LLMs and their evaluation scores. While we do not observe signs of memorization, we acknowledge that further controlled experiments are necessary to rigorously assess this risk.

Ethics Statement

All annotators involved in the collection of human judgments for the GUMSum dataset have given their consent to participate in the study and to allow us to publish the collected scores. They were paid according to national standards.

We used AI-assistants for improving the clarity of the text and of the figures.

Acknowledgments

We would like to thank the members of the MaiNLP lab for their valuable feedback, especially to Andreas Säuberli, Diego Frassinelli, Shijia Zhou, Soh-Eun (Ryan) Shim, and Domenico De Cristofaro. We also thank Pingjun Hong, Yuchen Mao, and Sebastian Loftus for their annotations. We recognize the support for Yang Janet Liu[7] and Barbara Plank through the ERC Consolidator Grant DIALECT 101043235.

[7]Work was carried out while at MaiNLP, LMU Munich.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Anthropic. 2024. Claude 3.5 sonnet. Accessed: 2025-02-09.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *Preprint*, arXiv:2307.15703.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sheila Castilho. 2020. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey. *Preprint*, arXiv:2006.14799.

Noam Dahan and Gabriel Stanovsky. 2025. The state and fate of summarization datasets: A survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7259–7278, Albuquerque, New Mexico. Association for Computational Linguistics.

Hang Dang and W. Bruce Croft. 2004. Overview of the trec 2004 robust retrieval track. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*, pages 1–10. National Institute of Standards and Technology (NIST).

Daniel Deutsch and Dan Roth. 2020. SacreROUGE: An open-source library for using and developing summarization evaluation metrics. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 120–125, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jessica Zosa Forde, Ruochen Zhang, Lintang Sutawika, Alham Fikri Aji, Samuel Cahyawijaya, Genta Indra Winata, Minghao Wu, Carsten Eickhoff, Stella Biderman, and Ellie Pavlick. 2024. Re-evaluating evaluation for multilingual summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19476–19493, Miami, Florida, USA. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77.

Théo Gigant, Camille Guinaudeau, Marc Decombas, and Frederic Dufaux. 2024. Mitigating the impact of reference quality on evaluation of summarization systems with reference-free metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19355–19368, Miami, Florida, USA. Association for Computational Linguistics.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, Volume 7 - 2024.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024. Benchmarking linguistic diversity of large language models. *Preprint*, arXiv:2412.10271.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jessica Lin and Amir Zeldes. 2025. GUM-SAGE: A novel dataset and approach for graded entity salience prediction. *Preprint*, arXiv:2504.10792.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Janet Liu and Amir Zeldes. 2023. GUMSum: Multi-genre data and evaluation for English abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9315–9327, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368, Singapore. Association for Computational Linguistics.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Lang. Resour. Eval.*, 52(1):101–148.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond n-grams: Rethinking evaluation metrics and strategies for multilingual abstractive summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19019–19035, Vienna, Austria. 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.

Ani Nenkova and Kathleen R. McKeown. 2011. Automatic Summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Keonwoong Noh, Seokjin Oh, and Woohwan Jung. 2024. Beyond reference: Evaluating high quality translations better than human references. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5111–5127, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. Accessed: 2025-02-09.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136, Sofia, Bulgaria. Association for Computational Linguistics.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Rezvaneh Rezapour, Sravana Reddy, Rosie Jones, and Ian Soboroff. 2022. What makes a good podcast summary? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2039–2046, New York, NY, USA. Association for Computing Machinery.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Michal Shur-Ofry, Bar Horowitz-Amsalem, Adir Rahamim, and Yonatan Belinkov. 2024. Growing a tail: Increasing output diversity in large language models. *Preprint*, arXiv:2411.02989.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Vivek Srivastava, Savita Bhat, and Niranjan Pedanekar. 2023. Hiding in plain sight: Insights into abstractive text summarization. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 67–74, Dubrovnik, Croatia. Association for Computational Linguistics.

Tianyi Tang, Hongyuan Lu, Yuchen Jiang, Haoyang Huang, Dongdong Zhang, Xin Zhao, Tom Kocmi, and Furu Wei. 2024. Not all metrics are guilty: Improving NLG evaluation by diversifying references. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6596–6610, Mexico City, Mexico. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Si Wu, John Wieting, and David A. Smith. 2025. Multiple references with meaningful variations improve literary machine translation. *Preprint*, arXiv:2412.18707.

Qwen Team: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A LLM-generated Summaries

For each source, we generate a single summary using each of the four LLMs. For SummEval, the sources corresponding to the 100 multi-reference summaries are taken into account. For DUC2004 Task 1, we generate summaries for the whole dataset (489 instances). For GUMSum, we focus on the dev and test set (in total 48 instances). Thus, we generate a total of 2548 summaries. We comply with the license of the existing datasets. For newly collected model outputs and human judgments, we follow the license of the corresponding underlying datasets. The codebase will be made publicly available upon publication.

A.1 Prompts

A.1.1 SummEval

`Article: {full_text}. Summarize the article in three sentences. Summary:`

A.1.2 DUC2004 Task 1

`The task is to create a very short single-document summary for the article below.`

`A very short summary should not be longer than 75 characters - this includes spaces and punctuation.`

`We will chop off characters beyond the 75th, so please do not include more than 75.`

`A very short summary could look like a newspaper headline, be a list of important terms or phrases separated by commas, a sentence, etc.`

`It should not contain any formatting, i.e., no indented lists, etc. Feel free to use your own words.`

`Article: {full_text} Summary:`

A.1.3 GUMSum

Following Liu and Zeldes (2023), a general prompt was used to instruct LLMs to generate summaries, as shown below. `Summarize the following article in 1 sentence. Make sure your summary is one sentence long and does not exceed 380 characters. Example of summary style: example`

`{doc_text}`

`Summary:`

A.2 LLM Output Evaluation

Table 3 reports the scores obtained by the four LLMs when using all available references for scoring.

B Reference-based Metrics

ROUGE. We use the `sacrerouge`[8] python implementation of ROUGE (Deutsch and Roth, 2020), with default parameters.

```
rouge = Rouge(
max_ngram = 4,
use_porter_stemmer = True,
remove_stopwords = False,
max_bytes = None,
max_words = None,
compute_rouge_l = True,
skip_bigram_gap_length = None,
scoring_function = "max", # or "average"
)
```

Notice that the implementation of ROUGE uses the Jackknife method when multiple references are provided.

BLEU. We use the `sacrebleu`[9] python implementation of BLEU (Post, 2018), with default parameters.

```
bleu = BLEU(
lowercase=False,
force=False,
tokenize=tokenize,
smooth_method='exp',
smooth_value=None,
effective_order=False) # True when used at
    the sentence level
)
```

METEOR. We use the Hugging Face version of Meteor, implemented through the `evaluate`[10] library, with default parameters, which wraps the NLTK implementation of the metric.[11]

```
nltk.translate.meteor_score.meteor_score(
    references: ~typing.Iterable[~typing.
    Iterable[str]], hypothesis: ~typing.Iterable
    [str], preprocess: ~typing.Callable[[str],
    str] = <method 'lower' of 'str' objects>,
    stemmer: ~nltk.stem.api.StemmerI = <
    PorterStemmer>, wordnet: ~nltk.corpus.reader
    .wordnet.WordNetCorpusReader = <
    WordNetCorpusReader in '/Users/stevenbird/
    nltk_data/corpora/wordnet'>, alpha: float =
    0.9, beta: float = 3.0, gamma: float = 0.5)
    -> float[source]
```

[8] https://github.com/danieldeutsch/sacrerouge
[9] https://github.com/mjpost/sacrebleu
[10] https://github.com/huggingface/evaluate
[11] https://www.nltk.org/api/nltk.translate.meteor_score.html

	SummEval							DUC							GUMSum						
	R-1	R-2	R-L	BLEU	MTR	BS	BLRT	R-1	R-2	R-L	BLEU	MTR	BS	BLRT	R-1	R-2	R-L	BLEU	MTR	BS	BLRT
Qwen	36.43	15.71	31.56	17.41	42.07	89.67	55.15	42.06	17.13	36.67	10.03	31.98	89.96	49.28	44.01	19.23	34.20	22.90	35.25	90.23	47.87
Llama3	37.50	18.09	32.79	22.07	46.09	90.04	56.61	37.44	12.74	32.79	6.45	24.97	89.87	46.01	43.51	19.96	35.34	24.65	36.70	90.33	48.50
Claude	36.03	17.80	31.89	20.84	46.75	89.83	56.07	47.69	20.92	41.73	12.86	39.48	91.27	56.94	44.99	18.65	34.51	20.85	37.62	90.40	50.13
GPT4	35.86	17.90	31.80	21.84	46.36	90.10	57.22	45.73	19.22	39.27	12.07	38.12	91.15	56.01	47.02	19.12	34.23	21.33	43.46	90.17	51.61

Table 3: **LLM scores on the multi-reference datasets.** *R*, *MTR*, *BS*, and *BLRT* are short for ROUGE, METEOR, BERTScore, and BLEURT respectively. All available references are used in the evaluation. For ROUGE and BLEURT, we consider the max-variation of the score.

BERTScore. We use the Hugging Face version of BERTScore, implemented through the `evaluate`[12] library, with default parameters, which wraps the `bert_score` implementation.[13] No TF-IDF weighting is used. Embeddings are obtained by using `FacebookAI/roberta-large`. We did not fine-tune the model. The corresponding hash is `roberta-large_L17_no-idf_version=0.3.12 (hug_trans=4.48.3)`.

BLEURT. We use the original Google version of BLEURT.[14] We used the recommended checkpoint,[15] which we did not fine-tune.

C Use of Reference-based Metrics

We survey papers published at ACL, EMNLP, and INLG from 2023 and 2024. Given the list of long and short accepted papers, we collected papers matching the keyword `summar*` in their title. After filtering out papers that are not about summarization research (e.g., *summarizing* the state of the art for a different topic), for each paper we checked whether one of our chosen metrics had been used for evaluation. Finally, we point out that some of the surveyed papers do not perform experimental work and thus do not evaluate model outputs (e.g., paper studying human evaluation); the reported percentages are thus slightly underestimated.

D Variability in Humans and LLMs

Figure 8 compares the variability observed in human-written summaries and in LLM-generated ones. To characterize LLM-generated summaries, given and instance i and a pair of human-written summaries for instance i R_{ji} and R_{zi}, we plot $P(o_i, o_j)$ where P is the lexical, syntactic, or semantic similarity. To characterize LLM-generated summaries, given and instance i system S_j and S_z, producing outputs o_j and o_z, we plot $P(o_{ji}, o_{zi})$.

When compared to summaries generated by different humans, those produced by various LLMs exhibit far less variation, particularly at the semantic and syntactic levels. Quantifying and mitigating the reduced richness and variability of LLM-generated content is an area of ongoing research (Guo et al., 2024; Shur-Ofry et al., 2024; Giulianelli et al., 2023) and, while not reflecting poor output quality, raises an open question about whether LLMs can fully replace human references.

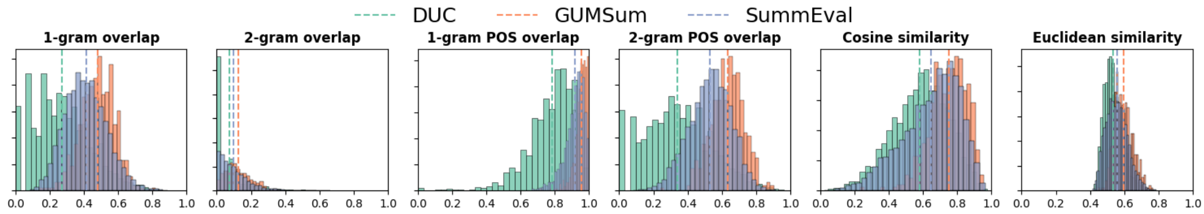
E Quality of Scraped References

One peculiarity of summarization datasets is that they are typically gathered from existing sources (Dahan and Stanovsky, 2025). This is the case of the CNN/DM dataset (Hermann et al., 2015; Nallapati et al., 2016), where articles are scraped from news websites, with the concatenated highlights acting as the target summary. The quality of such targets has been criticized in previous studies (Kryscinski et al., 2019; Srivastava et al., 2023), as they contain extraneous facts, references to other articles, and other issues.

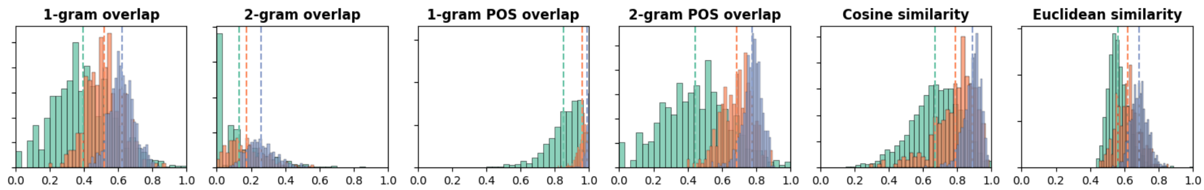
To better investigate the impact of these issues—and thus, quantify the gap between scraped summaries and crowd sourced ones—we leverage the additional references in SummEval. To this end, we treat the scraped CNN/DM references as hypotheses, and we score them against the 10 crowd-sourced ones in the SummEval dataset. We assume the latter provide a superior reference set for two main reasons: a) crowd-sourced references are specifically collected to act as summaries, with clear guidelines and an emphasis on quality; b) the larger cardinality of the crowd sourced human references (cf. Table 1) allows for a more comprehensive view of human-produced summaries. We compare these score to those of system outputs.

Figure 9 reports the scores of the original CN-

[12]https://github.com/huggingface/evaluate
[13]https://github.com/Tiiiger/bert_score
[14]https://github.com/google-research/bleurt
[15]available at https://storage.googleapis.com/bleurt-oss-21/BLEURT-20.zip



(a) human-written summaries



(b) LLM-generated summaries

Figure 8: Variation in human-produced summaries (top) and LLM-generated summaries (bottom).

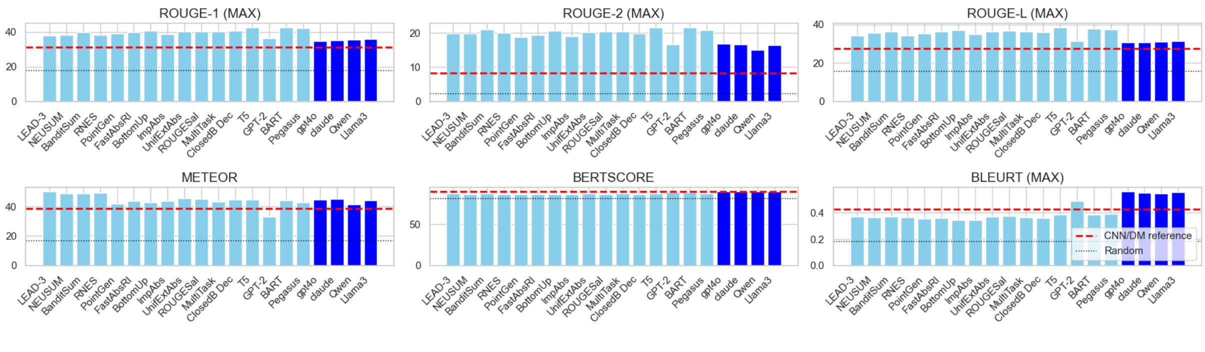


Figure 9: **Models and CNN/DM reference compared to** 10 **crowd-sourced references in SummEval.** SummEval models are in lighter blue, LLMs in darker blue. The red line indicates the CNN/DM reference score, and the black dotted line represents a random baseline using a different news article as the hypothesis.

N/DM references (red line) and those of the outputs of the systems (blue bars). We also report a random baseline (dotted line) in which a summary of a different document randomly sampled from the collection is used as hypothesis.

Notably, the original CNN/DM references perform *worse* than all model outputs in all cases but one when evaluated using n-gram-matching metrics. When using BERTScore, CNN/DM receives a score close to that of the outputs from LLMs. BLEURT rates the original reference higher than SummEval system outputs (except for GPT-2), but still lower than all LLM-generated outputs. These observations corroborate previous concerns on reference summary quality, especially when used to score high-quality outputs (Goyal et al., 2023).

F Instance-level variation

Figure 10 and Figure 11 contain the histogram of the instance-level variability for GUMSum and DUC respectively.

G Model Ranking

Table 4 contains the analysis of the rank stability when using one single reference ($k = 1$). For SummEval, we considering the cases of ranking the models studied by Fabbri et al. (2021), the LLMs and the combination of the two separately.

Figure 12 shows the Kendall tau correlation between ranks as the number k of references increases. We use the mean version of ROUGE.

H System-level Correlation

Figure 13 shows the instance-level correlation for SummEval (top) and GUMSum (bottom) using $\text{ROUGE}_{\text{avg}}$. Figure 14 show the system-level correlation on SummEval with $\text{ROUGE}_{\text{max}}$ (top) and $\text{ROUGE}_{\text{avg}}$ (bottom).

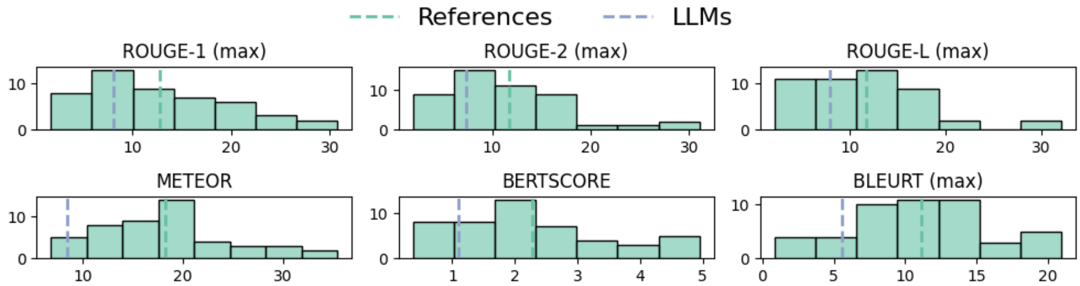


Figure 10: **Ranges of variability at the instance level on GUMSum.** For each instance, we compute the range of the scores of the references scored against the remaining ones.

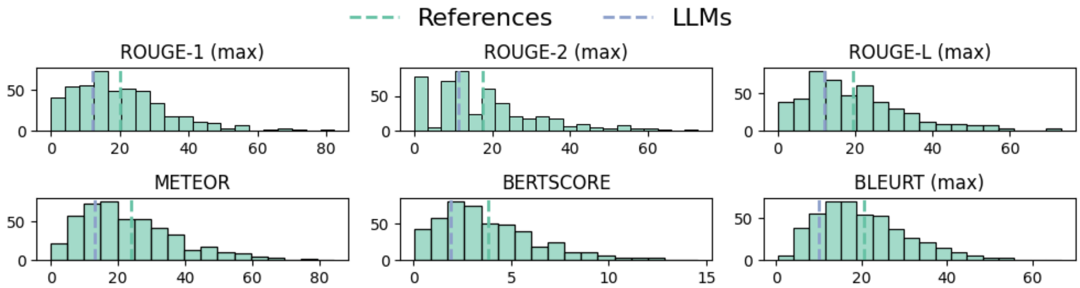


Figure 11: **Ranges of variability at the instance level on DUC.** For each instance, we compute the range of the scores of the references scored against the remaining ones.

		$R1_{mean}$	$R2_{mean}$	$R3_{mean}$	$R4_{mean}$	RL_{mean}	$R1_{max}$	$R2_{max}$	$R3_{max}$	$R4_{max}$	RL_{max}	BLEU	MTR	BS	BLRT
$\text{SummEval}_{\text{all models}}$	min	-.04	-.07	.05	-.09	.12	-.06	-.09	-.04	-.06	.04	-.27	.41	.54	.67
	avg	.51	.56	.57	.51	.63	.51	.56	.56	.52	.62	.38	.76	.79	.86
	std	.16	.14	.12	.14	.12	.18	.15	.13	.13	.14	.16	.07	.06	.04
	max	.91	.87	.91	.82	.92	.87	.86	.88	.85	.92	.82	.94	.95	.98
$\text{SummEval}_{\text{pre-LLMs}}$	min	-.05	-.03	-.05	-.15	0	-.05	-.03	-.12	-.28	.-07	-.30	.57	.28	.50
	avg	.48	.49	.48	.39	.55	.49	.50	.57	.39	.55	.39	.78	.68	.79
	std	.14	.13	.13	.15	.13	.13	.13	.15	.16	.12	.17	.06	.09	.07
	max	.87	.88	.90	.83	.90	.88	.85	.87	.88	.90	.85	.97	.93	.97
$\text{SummEval}_{\text{LLMs}}$	min	-.33	0	-.33	-1	0	-.33	-.33	.67	-1	0	-.67	0	-67	0
	avg	.59	.54	.49	.48	.59	.58	.52	.48	.49	.57	.59	.60	.62	.80
	std	.31	.32	.33	.35	.31	.32	.32	.34	.36	.31	.32	.31	.28	.21
	max	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GUMSum	min	0	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	0	-1	.33
	avg	.53	.42	.22	.15	.33	.51	.35	.14	.16	.35	.28	.56	.17	.86
	std	.32	.40	.44	.47	.41	.32	.43	.47	.47	.43	.45	.31	.48	.19
	max	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DUC	min	.67	.67	.33	-.33	.67	.67	.67	.33	-.33	.67	.33	.67	.33	.67
	avg	.99	.99	.94	.82	.98	.99	.97	.91	.78	.99	.88	.89	.71	.88
	std	.05	.05	.12	.22	.08	.05	0.9	.15	.25	.05	.17	.16	.24	.16
	max	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 4: **Rank stability with a single reference.** We ranked systems 100 times and compute the Kendall tau correlation among such rankings. *R*, *MTR*, *BS*, and *BLRT* are short for ROUGE, METEOR, BERTScore, and BLEURT respectively.

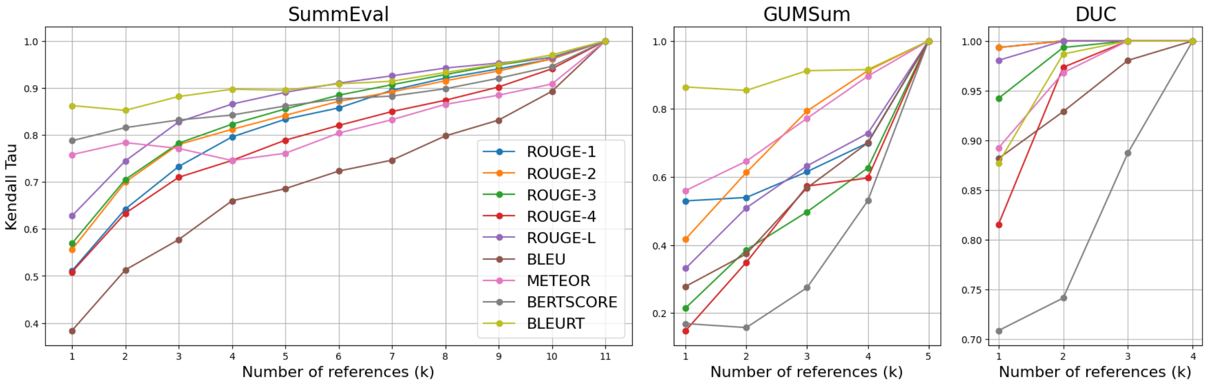


Figure 12: **Rank stability when increasing the number of references** over all three datasets. For ROUGE, we show the mean variant. Notice we use different ranges for the y axes for each dataset to improve readability.

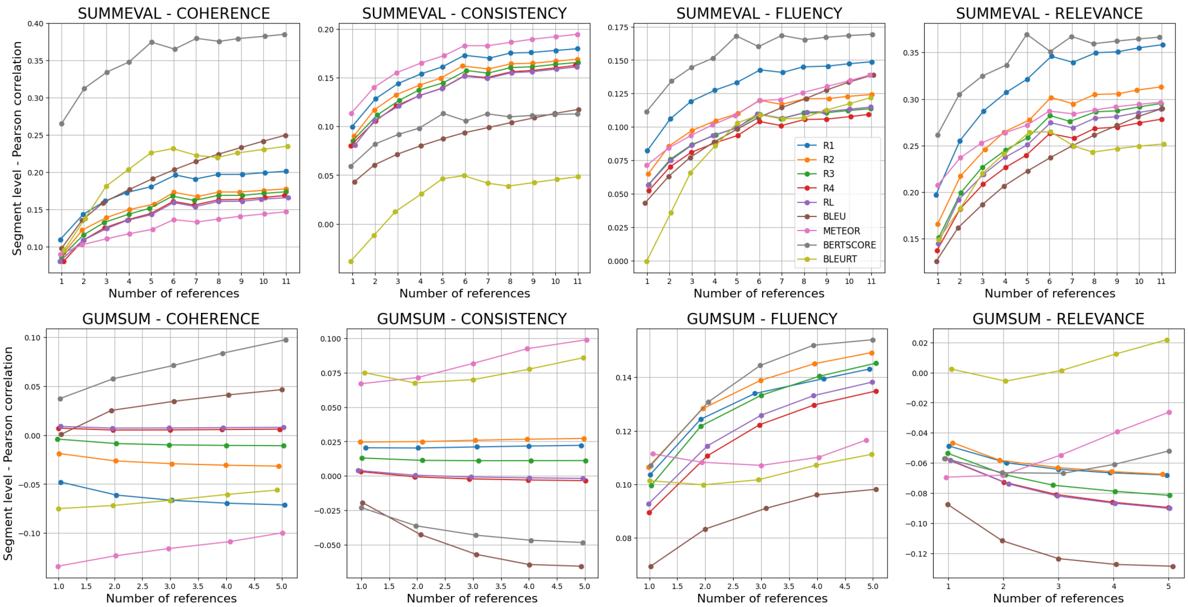


Figure 13: Pearson correlation at the instance level on SummEval (top) and GUMSum (bottom) using $\text{ROUGE}_{\text{mean}}$.

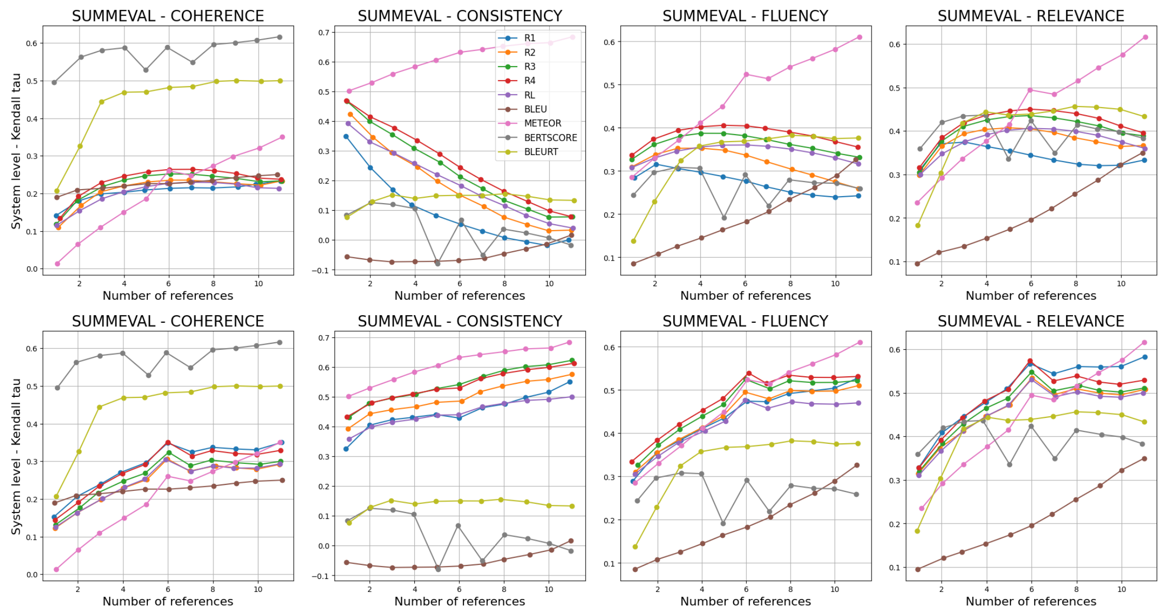


Figure 14: Kendall tau at the system level on SummEval using $\text{ROUGE}_{\text{max}}$ (top) and $\text{ROUGE}_{\text{mean}}$ (bottom).