Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation

Yixin Liu[*1] **Alexander R. Fabbri**[*2] **Pengfei Liu**[3] **Yilun Zhao**[1]
Linyong Nan[1] **Ruilin Han**[1] **Simeng Han**[1] **Shafiq Joty**[2]
Chien-Sheng Wu[2] **Caiming Xiong**[2] **Dragomir Radev**[1]
[1]Yale University, [2]Salesforce AI, [3]Carnegie Mellon University
yixin.liu@yale.edu, afabbri@salesforce.com

Abstract

Human evaluation is the foundation upon which the evaluation of both summarization systems and automatic metrics rests. However, existing human evaluation studies for summarization either exhibit a low inter-annotator agreement or have insufficient scale, and an in-depth analysis of human evaluation is lacking. Therefore, we address the shortcomings of existing summarization evaluation along the following axes: (1) We propose a modified summarization salience protocol, *Atomic Content Units* (ACUs), which is based on fine-grained semantic units and allows for a high inter-annotator agreement. (2) We curate the Robust Summarization Evaluation (**RoSE**) benchmark, a large human evaluation dataset consisting of 22,000 summary-level annotations over 28 top-performing systems on three datasets. (3) We conduct a comparative study of four human evaluation protocols, underscoring potential confounding factors in evaluation setups. (4) We evaluate 50 automatic metrics and their variants using the collected human annotations across evaluation protocols and demonstrate how our benchmark leads to more statistically stable and significant results. The metrics we benchmarked include recent methods based on large language models (LLMs), GPTScore and G-Eval. Furthermore, our findings have important implications for evaluating LLMs, as we show that LLMs adjusted by human feedback (e.g., GPT-3.5) may overfit unconstrained human evaluation, which is affected by the annotators' prior, input-agnostic preferences, calling for more robust, targeted evaluation methods.

1 Introduction

Human evaluation plays an essential role in both assessing the rapid development of summarization systems in recent years (Lewis et al., 2020a; Zhang et al., 2020a; Brown et al., 2020; Sanh et al., 2022; He et al., 2022) and in assessing the ability of automatic metrics to evaluate such systems as a proxy for manual evaluation (Bhandari et al., 2020; Fabbri et al., 2022a; Gao and Wan, 2022). However, while human evaluation is regarded as the gold standard for evaluating both summarization systems and automatic metrics, as suggested by Clark et al. (2021) an evaluation study does not become "gold" automatically without proper practices. For example, achieving a high inter-annotator agreement among annotators can be difficult (Goyal et al., 2022), and there can be a near-zero correlation between the annotations of crowd-workers and expert annotators (Fabbri et al., 2022a). Also, a human evaluation study without a large enough sample size can fail to find statistically significant results due to insufficient statistical power (Card et al., 2020).

Therefore, we believe it is important to ensure that **human evaluation can indeed serve as a solid foundation for evaluating summarization systems and automatic metrics**. For this, we propose using a robust ***human evaluation protocol*** for evaluating the salience of summaries that is more objective by dissecting the summaries into fine-grained content units and defining the annotation task based on those units. Specifically, we introduce the *Atomic Content Unit* (ACU) protocol for summary salience evaluation (§3), which is modified from the Pyramid (Nenkova and Passonneau, 2004) and LitePyramid (Shapira et al., 2019) protocols. We demonstrate that with the ACU protocol, a high inter-annotator agreement can be established among crowd-workers, which leads to more stable system evaluation results and better reproducibility.

We then collect, through both in-house annotation and crowdsourcing, **RoSE**, a large ***human evaluation benchmark*** of human-annotated summaries with the ACU evaluation protocol on recent state-of-the-art summarization systems, which yields higher statistical power (§4). To support evaluation across datasets and domains, our benchmark consists of test sets over three summarization datasets, CNN/DailyMail (CNNDM) (Nalla-

[*] Equal contribution

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4140–4170
July 9-14, 2023 ©2023 Association for Computational Linguistics

Statistical Power §4.1	– High statistical power is difficult to reach for human evaluation of similar-performing systems. – Increasing the sample size of human evaluation effectively raises statistical power.
Summary Length §4.2	– Summaries from different summarization systems show a large difference in average length. – Difference in summary length is not well-reflected by automatic evaluation metrics.
Evaluation Protocol Comparison §5.2	– Reference-free and reference-based human evaluation results have a near-zero correlation. – Reference-free human evaluation strongly correlates with input-agnostic, annotator preference. – Annotator's input-agnostic preference has a strong positive correlation with summary lengths. – Annotator's input-agnostic preference does not favor reference summaries. – Compared to smaller, fine-tuned models, zero-shot large language models (e.g. GPT-3) perform better under reference-free evaluation, but worse under reference-based evaluation.
Evaluating Automatic Metrics §6.1 & §6.2	– A higher-powered human evaluation dataset can lead to a more robust automatic metric evaluation, as shown by a tighter confidence interval and higher statistical power of metric evaluation. – Automatic metric performance differs greatly under different human evaluation protocols. – Automatic metrics show relatively strong system-level correlation and moderate summary-level correlation with our robust human evaluation protocol.

Table 1: Summary of the key findings in our work.

pati et al., 2016), XSum (Narayan et al., 2018), and SamSum (Gliwa et al., 2019), and annotations on the validation set of CNNDM to facilitate automatic metric training. To gain further insights into the characteristics of different evaluation protocols, we conduct ***human evaluation with three other protocols*** (§5). Specifically, we analyze protocol differences in the context of both fine-tuned models and large language models (LLMs) in a zero-shot setting such as GPT-3 (Brown et al., 2020). We find that different protocols can lead to drastically different results, which can be affected by annotators' prior preferences, highlighting the importance of aligning the protocol with the summary quality intended to be evaluated. We note that our benchmark enables a more trustworthy ***evaluation of automatic metrics*** (§6), as shown by statistical characteristics such as tighter confidence intervals and more statistically significant comparisons (§6.2). Our evaluation includes recent methods based on LLMs (Fu et al., 2023; Liu et al., 2023), and we found that they cannot outperform traditional metrics despite their successes on related benchmarks such as SummEval (Fabbri et al., 2022a).

We summarize our key findings in Tab. 1. Our contributions are the following: (1) We propose the ACU protocol for high-agreement human evaluation of summary salience. (2) We curate the **RoSE** benchmark, consisting of 22000 summary-level annotations and requiring over 150 hours of in-house annotation, across three summarization datasets, which can lay a solid foundation for training and evaluating automatic metrics.[1] (3) We compare four human evaluation protocols for summarization and show how they can lead to drastically different model preferences. (4) We evaluate automatic metrics across different human evaluation protocols and call for human evaluation to be conducted with a clear evaluation target aligned with the evaluated systems or metrics, such that *task-specific* qualities can be evaluated without the impact of general, *input-agnostic* preferences of annotators. We note that the implications of our findings can become even more critical with the progress of LLMs trained with human preference feedback (Ouyang et al., 2022) and call for a more rigorous human evaluation of LLM performance.

[1]We release our benchmark and evaluation scripts at https://github.com/Yale-LILY/ROSE.

2 Related Work

Human Evaluation Benchmarks Human annotations are essential to the analysis of summarization research progress. Thus, recent efforts have focused on aggregating model outputs and annotating them according to specific quality dimensions (Huang et al., 2020; Bhandari et al., 2020; Stiennon et al., 2020; Zhang and Bansal, 2021; Fabbri et al., 2022a; Gao and Wan, 2022). The most relevant work to ours is Bhandari et al. (2020), which annotates summaries according to semantic content units, motivated by the Pyramid (Nenkova and Passonneau, 2004) and LitePyramid (Shapira et al., 2019) protocols. However, this benchmark only covers a single dataset (CNNDM) without a focus on similarly-performing state-of-the-art systems, which may skew metric analysis (Tang et al., 2022a) and not fully reflect realistic scenarios (Deutsch et al., 2022). In contrast, our benchmark consists only of outputs from recently-introduced models over three datasets.

Summarization Meta-Evaluation With a human evaluation dataset, there exist many directions of meta-evaluation, or re-evaluation of the current state of evaluation, such as metric performance analyses, understanding model strengths, and human evaluation protocol comparisons.

Within metric meta-analysis, several studies have focused on the analysis of ROUGE (Lin, 2004b), and its variations (Rankel et al., 2013; Graham, 2015), across domains such as news (Lin, 2004a), meeting summarization (Liu and Liu, 2008), and scientific articles (Cohan and Goharian, 2016). Other studies analyze a broader set of metrics (Peyrard, 2019; Bhandari et al., 2020; Deutsch and Roth, 2020; Fabbri et al., 2022a; Gabriel et al., 2021; Kasai et al., 2022b), including those specific to factual consistency evaluation (Kryscinski et al., 2020; Durmus et al., 2020; Wang et al., 2020; Maynez et al., 2020; Laban et al., 20d; Fabbri et al., 2022b; Honovich et al., 2022; Tam et al., 2022).

Regarding re-evaluating model performance, a recent line of work has focused on evaluating zero-shot large language models (Goyal et al., 2022; Liang et al., 2022; Tam et al., 2022), noting their high performance compared to smaller models.

As for the further understanding of human evaluation, prior work has compared approaches to human evaluation (Hardy et al., 2019), studied annotation protocols for quality dimensions such as linguistic quality (Steen and Markert, 2021) and factual consistency (Tang et al., 2022b), and noted the effects of human annotation inconsistencies on system rankings (Owczarzak et al., 2012). The unreliability and cost of human evaluation in certain settings have been emphasized (Chaganty et al., 2018; Clark et al., 2021), with some work noting that thousands of costly data points may need to be collected in order to draw statistically significant conclusions (Wei and Jia, 2021). Our meta-analysis focuses on this latter aspect, and we further analyze potential confounding factors in evaluation such as length and protocol design, with respect to both small and large zero-shot language models.

3 Atomic Content Units for Summarization Evaluation

We now describe our Atomic Content Unit (ACU) annotation protocol for reference-based summary salience evaluation, including the procedure of writing ACUs based on reference summaries and matching the written ACUs with system outputs.

3.1 Preliminaries

In this work, we focus on a specific summarization meta-evaluation study on *summary salience*. Salience is a desired summary quality that requires the summary to include all and only important information of the input article. The human evaluation of summary salience can be conducted in either *reference-free* or *reference-based* manners. The former asks the annotators to assess the summary directly based on the input article (Fabbri et al., 2022a), while the latter requires the annotators to assess the information overlap between the system output and reference summary (Bhandari et al., 2020), under the assumption that the reference summary is the gold standard of summary salience.[2] Given that reference-based protocols are more constrained, we focus on ***reference-based evaluation*** for our human judgment dataset collection, and we conduct a comparison of protocols in §5.

3.2 ACU Annotation Protocol

Inspired by the Pyramid (Nenkova and Passonneau, 2004) and LitePyramid (Shapira et al., 2019) protocols and subsequent annotation collection efforts (Bhandari et al., 2020; Zhang and Bansal, 2021), the ACU protocol is designed to reduce the subjectivity of reference-based human evaluation by simplifying the basic annotation unit – the annotators only need to decide on the presence of a single fact, extracted from one text sequence, in another text sequence, to which a binary label can be assigned with more objectivity. Specifically, the evaluation process is decomposed into two steps: (1) **ACU Writing** – extracting facts from one text sequence, and (2) **ACU Matching** – checking for the presence of the extracted facts in another sequence. We formulate the ACU protocol as a ***recall-based*** protocol, such that the first step only needs to be performed once for the reference summary, allowing for reproducibility and reuse of these units when performing matching on new system outputs.

ACU Writing While the LitePyramid approach defines its basic content unit as a sentence containing a brief fact, we follow Bhandari et al. (2020) to emphasize a shorter, more fine-grained information unit. Specifically, we define the ACU protocol with the concept of ***atomic facts*** – elementary information units in the reference summaries, which no

[2]We note salience can be an inherently subjective quality, and the reference summary of common datasets may not always be the actual "gold standard," discussed more in §7.

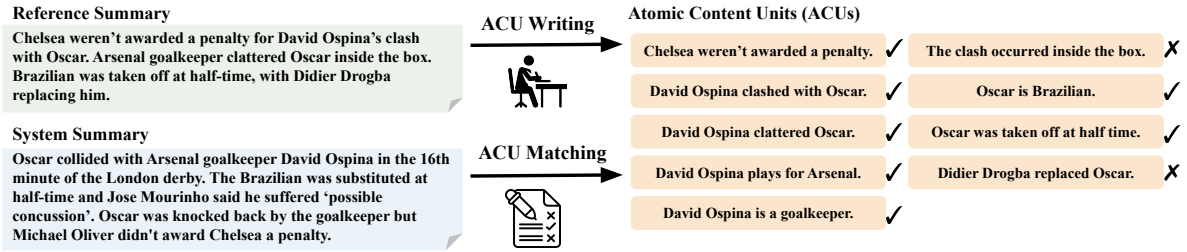


Figure 1: Example of a reference summary, a system summary and corresponding ACU annotations on CNNDM.

longer need to be further split for the purpose of reducing ambiguity in human evaluation.[3] Then, ACUs are constructed based on one atomic fact and other minimal, necessary information.

Fig. 1 shows an example of the written ACUs. To ensure annotation quality, we (the authors) write all the ACUs used in this work. We define guidelines to standardize the annotation process; for each summary sentence the annotator creates an ACU constituting the main information from the subject of the main clause (e.g., root), followed by additional ACUs for other facts while including the minimal necessary information from the root. We provide rules for dealing with quotations, extraneous adjectives, noisy summaries, and additional cases. We note that there can still be inherent subjectivity in the written ACUs among different annotators even with the provided guidelines. However, such subjectivity should be unbiased in summary comparison because all the candidate summaries are evaluated by the same set of written ACUs.

ACU Matching Given ACUs written for a set of reference summaries, our protocol evaluates summarization system performance by checking the presence of the ACUs in the system-generated summaries as illustrated in Fig. 1. For this step, we recruit annotators on Amazon Mechanical Turk[4] (MTurk). The annotators must pass a qualification test, and additional requirements are specified in Appendix A. Besides displaying the ACUs and the system outputs, we also provide the reference summaries to be used as context for the ACUs.

Scoring Summaries with ACU ACU matching annotations can be aggregated into summary scores. We first define an un-normalized ACU score f of a candidate summary s given a set of ACUs $\mathcal{A}$ as:

$$f(s, \mathcal{A}) = \frac{|\mathcal{A}_s|}{|\mathcal{A}|}, \quad (1)$$

[3]We note that it can be impossible to provide a practical definition of atomic facts. Instead, we use it as a general concept for fine-grained information units.

[4]https://www.mturk.com/

Dataset	Split	#Doc.	#Sys.	#ACU	#Summ.
CNNDM	Test	500	12	5.6k	6k
CNNDM	Valid	1,000	8	11.6k	8k
XSum	Test	500	8	2.3k	4k
SamSum	Test	500	8	2.3k	4k

Table 2: Statistics of the collected annotations. **#Doc.** is the number of input documents, **#Sys.** is the number of summarization systems used for collection. **#ACU** is the total number of written ACUs. **#Summ.** is the total number of summary-level annotations, which are aggregated over three annotations on the test sets, and a single annotation on the validation set of CNNDM.

where $\mathcal{A}_s$ is a subset of $\mathcal{A}$ that is matched with s. We note that f by default is a *recall* based score with respect to the reference summary r. Therefore, we also define a *normalized* ACU score $\tilde{f}$ as:

$$\tilde{f}_\alpha(s, \mathcal{A}, r) = e^{\min\left(0, \frac{1-\frac{|s|}{|r|}}{\alpha}\right)} f(s, \mathcal{A}), \quad (2)$$

where $|s|$, $|r|$ are the length (i.e., number of words) of the candidate summary s and the reference summary r respectively, and α is a positive number controlling the strength of the normalization. This normalization is in effect a *redundancy penalty*, which penalizes the summaries longer than the reference and resembles the brevity penalty in BLEU (Papineni et al., 2002). In practice, we set the value of α by de-correlating $\tilde{f}$ with the summary length using the collected ACU annotations.

3.3 ACU Annotation Collection

We collect ACU annotations on three summarization datasets: CNNDM (Nallapati et al., 2016), XSum (Narayan et al., 2018), and SamSum (Gliwa et al., 2019). To reflect the latest progress in text summarization, we collect and annotate the generated summaries of pre-trained summarization systems proposed in recent years.[5] Detailed informa-

[5]We release all of the system outputs with a unified, cased, untokenized format to facilitate future research.

tion about the summarization systems we used can be found in Appendix A.2.

Table 2 shows the statistics of the collected annotations. The annotations are collected from the test set of the above datasets, and additionally from the validation set of CNNDM to facilitate the training of automatic evaluation metrics. In total, we collect around 21.8k ACU-level annotations and around 22k summary-level annotations, aggregated over around 50k individual summary-level judgments.

To calculate inter-annotator agreement, we use Krippendorff's alpha (Krippendorff, 2011). The aggregated summary-level agreement score of ACU matching is 0.7571, and the ACU-level agreement score is 0.7528. These agreement scores are higher than prior collections, such as RealSumm (Bhandari et al., 2020) and SummEval (Fabbri et al., 2022a), which have an average agreement score of crowd-workers 0.66 and 0.49, respectively.

4 RoSE Benchmark Analysis

We first analyze the robustness of our collected annotations and a case study on the system outputs.

4.1 Power Analysis

We analyze the *statistical power* of our collected human annotations to study whether it can yield stable and trustworthy results (Card et al., 2020). Statistical power is the probability that the null hypothesis of a statistical significance test is rejected given there is a real effect. For example, for a human evaluation study that compares the performance of two genuinely different systems, a statistical power of 0.80 means there is an 80% chance that a significant difference will be observed. Further details can be found in Appendix B.1.

We conduct the power analysis for *pair-wise* system comparisons with ACU scores (Eq. 1) focusing on two factors, the *number of test examples* and the *observed system difference*. Specifically, we run the power analysis with varying sample sizes, and group the system pairs into buckets according to their performance difference, as determined by ROUGE1 recall scores (Fig.2).[6] We observe the following: (1) **A high statistical power[7] is difficult to reach when the system performance is similar.**

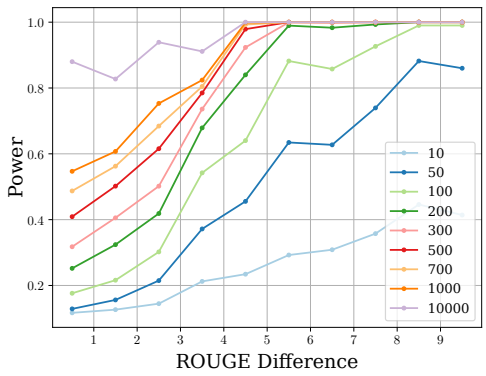


Figure 2: **Power analysis** of human evaluation for system comparison on the annotated CNNDM test examples. Different lines represent results with different sample sizes. The system pairs are grouped by performance differences in ROUGE1 recall scores.

Notably, while the sample size of the human evaluation performed in recent work is typically around 50-100,[8] such sample size can only reach a power of 0.80 when the ROUGE1 recall score difference is above 5. (2) **Increasing the sample size can effectively raise the statistical power.** For example, when the system performance difference is within the range of 1-2 points, the power of a 500-sample set is around 0.50 while a 100-sample set only has a power of around 0.20. The results of power analysis on three datasets with both ROUGE and ACU score differences are provided in Appendix B.2 with the same patterns, which indicates that our dataset can provide more stable summarization system evaluation thanks to its higher statistical power.

4.2 Summarization System Analysis

As a case study, in Tab. 3 we analyze the summary characteristics of the recent summarization systems we collected on the CNNDM test set. XSum and SamSum results are shown in Appendix A.3. Apart from the ACU scores, we note that **the average summary length of different systems can greatly vary**, and such differences are not always captured by the widely-used ROUGE F1. For example, the length of GSum (Dou et al., 2021) is around 40% longer than GLOBAL (Ma et al., 2021) while they have very similar ROUGE1 F1 scores. Besides, we note **all systems in Tab. 3 have longer summaries than the reference summaries**, whose average length is only 54.93. This can be a potential risk to users who may prefer shorter, more concise

[6] We note that these scores are proxies of the true system differences, and the power analysis is based on the assumption that the systems have significantly different performance.

[7] An experiment is usually considered sufficiently powered if the statistical power is over 0.80.

[8] We provide a brief survey of the practices of human evaluation in recent text summarization research in Appendix F.

System	ACU	*n*ACU	Len	R1F
GSum (Dou et al., 2021)	44.47	34.87	77.61	45.47
MatchSum (Zhong et al., 2020)	42.50	33.69	74.99	43.84
BRIO-Ext (Liu et al., 2022)	41.72	33.58	73.67	44.44
BART (Lewis et al., 2020a)	38.83	32.34	71.00	44.04
CTRLSum (He et al., 2020)	**44.58**	36.13	70.56	45.69
BRIO (Liu et al., 2022)	44.03	**37.20**	69.58	**47.83**
CLIFF (Cao and Wang, 2021)	38.51	32.96	67.74	44.19
PEGASUS (Zhang et al., 2020a)	37.56	32.03	65.65	43.80
SimCLS (Liu and Liu, 2021)	40.47	36.01	62.91	46.46
FROST (Narayan et al., 20d)	38.44	33.68	62.65	44.90
GOLD (Pang and He, 2021)	38.10	33.80	60.65	44.86
GLOBAL (Ma et al., 2021)	36.40	34.07	55.50	45.17

Table 3: Summarization system analysis on CNNDM. **ACU** is the ACU score (Eq. 1), ***n*ACU** is the normalized ACU score (Eq. 2), **Len** is the average summary length, and **R1F** is the ROUGE1 F1 score. **ACU** and ***n*ACU** are calculated on the 500 annotated examples (the value is multiplied by 100) while **Len** and **R1F** are calculated on the entire test set. The systems are sorted by **Len**.

summaries. Meanwhile, the systems that generate longer summaries may be favored by users who prefer more informative summaries. Therefore, we join the previous work (Sun et al., 2019; Song et al., 2021; Gehrmann et al., 2022; Goyal et al., 2022) in advocating **treating summary lengths as a separate aspect of summary quality in evaluation**, as in earlier work in summarization research.[9]

5 Evaluating Annotation Protocols

Apart from ACU annotations, we collect human annotations with three different protocols to better understand their characteristics. Specifically, two reference-free protocols are investigated: *Prior* protocol evaluates the annotators' preferences of summaries *without* the input document, while *Ref-free* protocol evaluates if summaries cover the salient information of the input document. We also consider one reference-based protocol, *Ref-based*, which evaluates the content similarity between the generated and reference summaries. Appendix D.1 provides detailed instructions for each protocol.

5.1 Annotation Collection

We collected three annotations per summary on a 100-example subset of the above CNNDM test set using the same pool of workers from our ACU qualification. Except for ACU, all of the summaries from different systems are evaluated within a single task with a score from 1 (worst) to 5 (best), similar to the EASL protocol (Sakaguchi and Van Durme, 2018). We collect (1) annotations of the 12 above systems, with an inter-annotator agreement (Krippendorff's alpha) of 0.3455, 0.2201, 0.2741 on *Prior*, *Ref-free*, *Ref-based* protocols respectively; (2) annotations for summaries from GPT-3 (Brown et al., 2020),[10] T0 (Sanh et al., 2022), BRIO, and BART to better understand annotation protocols with respect to recently introduced large language models applied to zero-shot summarization.

	Prior	Ref-free	Ref-based	*n*ACU
Prior	-	**0.926**	-0.061	0.048
Ref-free	0.926	-	**-0.247**	-0.093
Ref-based	-0.061	-0.247	-	0.762
*n*ACU	0.048	-0.093	0.762	-
Len.	0.833	0.875	-0.550	-0.296

Table 4: *System-level* Pearson's correlations between different protocols on the fine-tuned models. ***n*ACU** is the normalized ACU score. **Len.** is the summary length.

5.2 Results Analysis

We investigate both the summary-level and system-level correlations of evaluation results of different protocols to study their inherent similarity. Details of correlation calculation are in Appendix C.

Results on Fine-tuned Models We show the system-level protocol correlation when evaluating the fine-tuned models in Tab. 4, and the summary-level correlation can be found in Appendix D.2. We use the *normalized* ACU score (Eq. 2) because the other evaluation protocols are supposed to resemble an F1 score, while the ACU score is by definition recall-based. We have the following observations:
(1) The *Ref-free* protocol has a strong correlation with the *Prior* protocol, suggesting that the latter may have a large impact on the annotator's document-based judgments.
(2) Both the *Prior* and *Ref-free* protocols have a strong correlation with summary length, showing that annotators may favor longer summaries.
(3) The *Ref-free* protocol and the *Ref-based* protocol have a negative correlation while ideally they are supposed to measure similar quality aspects.

We perform power analysis on the results following the procedure in §4.1 and found that ACU protocol can yield higher statistical power than the *Ref-based* protocol, suggesting that the ACU protocol leads to more robust evaluation results. We also found that the reference-free *Prior* and *Ref-free*

[9] For example, the DUC evaluation campaigns set a pre-specified maximum summary length, or summary budget.

[10] We use the "text-davinci-002" version of GPT-3.

	Prior	Ref-free	Ref-based	ACU	Len.
BART	3.58	3.52	2.93	0.367	69.5
BRIO	3.51	3.49	**3.07**	**0.429**	66.4
T0	3.33	3.24	2.84	0.295	61.6
GPT-3	**3.72**	**3.76**	2.74	0.268	69.5
Ref.	2.85	2.94	-	-	54.9

Table 5: Model performance under different annotation protocols. **Len.** is the summary length. **Ref.** is the reference summary. *Prior*, *Ref-free*, *Ref-based* protocols have a score range from 1 to 5.

protocols have higher power than the reference-based protocols. However, we note that they are not directly comparable because they have different underlying evaluation targets, as shown by the near-zero correlation between them. Further details are provided in Appendix D.2.

Results on Large Language Models The results are shown in Tab. 5. Apart from the system outputs, we also annotate reference summaries for reference-free protocols. We found that **under the *Ref-free* protocol, GPT-3 receives the highest score while the reference summary is the least favorite one**, similar to the findings of recent work (Goyal et al., 2022; Liang et al., 2022). However, we found the same pattern with the *Prior* protocol, showing that the **annotators have a *prior* preference for GPT-3**. We provide an example in Appendix D.2 comparing GPT-3 and BRIO summaries under different protocols. Given the strong correlation between the *Prior* and *Ref-free* protocols, we note that there is a risk that the annotators' decisions are affected by their prior preferences that are not genuinely related to the task requirement. As a further investigation, we conduct an annotator-based case study including 4 annotators who annotated around 20 examples in this task, in which we compare two summary-level correlations (Eq. 3) given a specific annotator: (1) the correlation between their own *Ref-free* protocol scores and *Prior* scores; (2) the correlation between their *Ref-free* scores and the *Ref-free* scores averaged over the other annotations on each example. We found that the average value of the former is 0.404 while the latter is only 0.188, suggesting that **the annotators' own *Prior* score is a better prediction of their *Ref-free* score than the *Ref-free* score of other annotators**.

6 Evaluating Automatic Metrics

We analyze several representative automatic metrics, with additional results in Appendix E on 50

	CNNDM		XSum		SamSum	
	Sys.	Sum.	Sys.	Sum.	Sys.	Sum.
ROUGE1	.788	**.468**	**.714**	**.293**	.929	.439
ROUGE2	.758	.453	.643	.266	**1.00**	.395
ROUGEL	**.879**	.454	.643	.258	.929	.415
METEOR	.758	.407	.571	.268	.857	.373
CHRF	.758	.436	.571	.275	.857	.396
BERTScore	.515	.448	.571	.277	.857	.417
BARTScore	.727	.453	**.714**	.282	.929	.430
QAEval	.849	.358	.429	.198	.929	.384
SummaQA	.727	.119	.143	.019	.643	.102
Lite^{3}Pyramid	.849	.452	**.714**	.245	**1.00**	**.467**
GPTScore	.636	.129	.214	.099	.429	.158
G-Eval-3.5	.412	.164	.429	.136	.857	.248
G-Eval-3.5-S	.364	.171	.429	.144	.857	.262
G-Eval-4	.779	.274	.691	.185	.929	.405

Table 6: The Kendall's correlation between the automatic metric scores and ACU scores of system outputs on CNNDM, XSum, and SamSum datasets. The correlation is calculated at both the system level and the summary level. We use the *recall* score of the automatic metrics when available to align with the ACU scores.

automatic metric variants. We focus the metric evaluation on ACU annotations because of two insights from §5: (1) Reference-based metrics should be evaluated with reference-based human evaluation. (2) ACU protocol provides higher statistical power than the summary-level *Ref-based* protocol.

6.1 Metric Evaluation with ACU Annotations

We use the correlations between automatic metric scores and ACU annotation scores of system outputs to analyze and compare automatic metric performance. The following metrics are evaluated:

(1) lexical overlap based metrics, **ROUGE** (Lin, 2004b), **METEOR** (Lavie and Agarwal, 2007), **CHRF** (Popović, 2015); (2) pre-trained language model based metrics, **BERTScore** (Zhang et al., 2020c), **BARTScore** (Yuan et al., 2021); (3) question-answering based metrics, **SummaQA** (Scialom et al., 2019), **QAEval** (Deutsch et al., 2021a); (4) **Lite^{3}Pyramid** (Zhang and Bansal, 2021), which automates the LitePyramid evaluation process; (5) evaluation methods based on large language models, **GPTScore** (Fu et al., 2023) and **G-Eval** (Liu et al., 2023), with two variants that are based on GPT-3.5[11] (**G-Eval-3.5**) and GPT-4[12] (OpenAI, 2023) (**G-Eval-4**) respectively. We note that for LLM-based evaluation we require the metric to calculate the *recall* score. For G-

[11]OpenAI's `gpt-3.5-turbo-0301`: https://platform.openai.com/docs/models/gpt-3-5.

[12]OpenAI's `gpt-4-0314`: https://platform.openai.com/docs/models/gpt-4.

Bucket	1	2	3	4	5	6
ROUGE1	.091	.636	1.00	1.00	1.00	1.00
ROUGE2	-.091	.818	.818	1.00	1.00	1.00
ROUGEL	.455	.818	1.00	1.00	1.00	1.00
METEOR	.091	.818	.818	.818	1.00	1.00
CHRF	.091	.818	.818	.818	1.00	1.00
BERTScore	-.091	.636	.636	.091	.818	1.00
BARTScore	-.091	.818	.818	.818	1.00	1.00
QAEval	.455	.818	1.00	.818	1.00	1.00
SummaQA	.636	.818	.636	.273	1.00	1.00
Lite^{3}Pyramid	.273	.818	1.00	1.00	1.00	1.00
GPTScore	.273	.091	.455	1.00	1.00	1.00
G-Eval-3.5	-.091	-.273	-.091	.818	1.00	1.00
G-Eval-3.5-S	-.091	-.273	-.273	.818	1.00	1.00
G-Eval-4	.091	.818	.636	1.00	1.00	1.00

Table 7: The *system-level* Kendall's correlation between the automatic metric and ACU scores on different *system pairs* grouped by their ACU score differences on the CNNDM dataset, into six equal-sized buckets. We use the *recall* score of the automatic metrics when available.

Eval-3.5 we report two variants that are based on greedy decoding (G-Eval-3.5) and sampling (G-Eval-3.5-S) respectively, Details of the LLM-based evaluation are in Appendix E.2.

Tab. 6 shows the results, with additional results of more metrics in Appendix E.3. We note:
(1) Several automatic metrics from the different families of methods (e.g., ROUGE, BARTScore) are all able to achieve a relatively high correlation with the ACU scores, especially at the system level.
(2) Metric performance varies across different datasets. In particular, metrics tend to have stronger correlations on the SamSum dataset and weaker correlations on the XSum dataset. We hypothesize that one reason is that the reference summaries of the XSum dataset contain more complex structures.
(3) Despite their successes (Fu et al., 2023; Liu et al., 2023) in other human evaluation benchmarks such as SummEval, LLM-based automatic evaluation cannot outperform traditional methods such as ROUGE on RoSE. Moreover, their low summary-level correlation with ACU scores suggests that their predicted scores may not be well-calibrated.

Following Deutsch et al. (2022), we further investigate metric performance when evaluating system pairs with varying performance differences. Specifically, we group the system pairs based on the difference of their ACU scores into different buckets and calculate the modified Kendall's correlation (Deutsch et al., 2022) on each bucket. The system pairs in each bucket are provided in Appendix E.4. Tab. 7 shows that **the automatic metrics generally perform worse when they are used to evaluate similar-performing systems.**

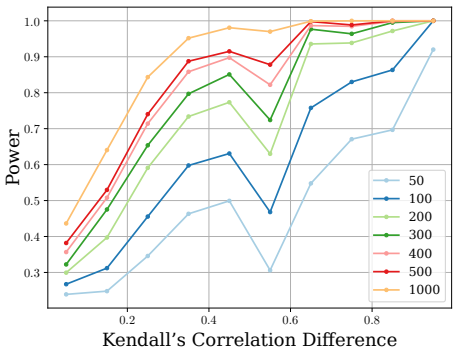


Figure 3: Power analysis of pair-wise metric comparison w.r.t. their *system-level* Kendall's correlation coefficients with ACU scores on CNNDM. The metric pairs are grouped by the correlation differences with ACU scores. Different lines represent different sample sizes.

6.2 Analysis of Metric Evaluation

We analyze the metric evaluation with respect to the statistical characteristics and the impact of different human evaluation protocols on metric evaluation.
Confidence Interval We select several representative automatic metrics and calculate the confidence intervals of their system-level correlations with the ACU scores using bootstrapping. Similar to Deutsch et al. (2021b), we find that the confidence intervals are large. However, we found that **having a larger sample size can effectively reduce the confidence interval**, which further shows the importance of increasing the *statistical power* of the human evaluation dataset as discussed in §4.1. We provide further details in Appendix E.5.
Power Analysis of Metric Comparison We conduct a power analysis of *pair-wise* metric comparison with around 200 pairs, which corresponds to the chance of a statistical significance result being found. More details can be found in Appendix E.6. The results are in Fig.3, showing similar patterns as in the power analysis of summarization system comparison (§4.1):
(1) Significant results are difficult to find when the metric performance is similar;
(2) Increasing the sample size can effectively increase the chance of finding significant results.
Correlations under Different Human Evaluation Protocols We analyze the metric correlations under different human evaluation protocols (§5). The results are shown in Tab. 8, with more results in Appendix E.7. We note: (1) Metric performance differs greatly under different protocols, likely because the protocols can have weak correlations with each other (§5.2). (2) The reference-based

Protocol	Prior	Ref-free	Ref-based	*n*ACU
ROUGE1	-0.061	-0.212	0.840	0.636
ROUGE2	0.000	-0.151	0.595	0.636
ROUGEl	-0.061	-0.212	0.779	0.636
METEOR	0.394	0.242	0.382	0.485
CHRF	0.576	0.424	0.199	0.485
BERTScore	-0.091	-0.182	0.779	0.485
BARTScore	-0.091	-0.182	0.656	0.364
QAEval	0.485	0.515	-0.076	0.151
SummaQA	0.515	0.424	0.260	0.303
Lite^{3}Pyramid	0.576	0.667	-0.168	0.121

Table 8: The *system-level* Kendall's correlation between the automatic metric and different human evaluation protocols on CNNDM dataset. We use the *F1* score of the automatic metrics when available.

automatic metrics generally perform better under reference-based evaluation protocols, but can have negative correlations with reference-free protocols.

7 Conclusion and Implications

We introduce RoSE, a benchmark whose underlying protocol and scale allow for more robust summarization evaluation across three datasets. With our benchmark, we re-evaluate the current state of human evaluation and its implications for both summarization system and automatic metric development, and we suggest the following:

(1) **Alignment in metric evaluation**. To evaluate automatic metrics, it is important to use an appropriate human evaluation protocol that captures the intended quality dimension to be measured. For example, *reference-based* automatic metrics should be evaluated by *reference-based* human evaluation, which disentangles metric performance from the impact of reference summaries.

(2) **Alignment in system evaluation**. We advocate for ***targeted evaluation***, which clearly defines the intended evaluation quality. Specifically, text summarization, as a conditional generation task, should be defined by both the source and target texts along with pre-specified, desired characteristics. Clearly specifying characteristics to be measured can lead to more reliable and objective evaluation results. This will be even more important for LLMs pretrained with human preference feedback for disentangling annotators' *prior* preferences for LLMs with the *task-specific* summary quality.

(3) **Alignment between NLP datasets and tasks**. Human judgments for summary quality can be diverse and affected by various factors such as summary lengths, and reference summaries are not always favored. Therefore, existing summarization datasets (e.g. CNNDM) should *only* be used for the appropriate tasks. For example, they can be used to define a summarization task with specific requirements (e.g. maximum summary lengths), and be important for studying reference-based metrics.

8 Limitations

Biases may be present in the data annotator as well as in the data the models were pretrained on. Furthermore, we only include English-language data in our benchmark and analysis. Recent work has noted that language models may be susceptible to learning such data biases (Lucy and Bamman, 2021), thus we request that the users be aware of potential issues in downstream use cases.

As described in Appendix D.1, we take measures to ensure a high quality benchmark. There will inevitably be noise in the dataset collection process, either in the ACU writing or matching step, and high agreement of annotations does not necessarily coincide with correctness. However, we believe that the steps taken to spot check ACU writing and filter workers for ACU matching allow us to curate a high-quality benchmark. Furthermore, we encourage the community to analyze and improve RoSE in the spirit of evolving, living benchmarks (Gehrmann et al., 2021).

For reference-based evaluation, questions about reference quality arise naturally. We also note that the original Pyramid protocol was designed for multi-reference evaluation and weighting of semantic content units, while we do not weight ACUs during aggregation. As discussed above, we argue that our benchmark and analysis are still valuable given the purpose of studying conditional generation and evaluating automatic metrics for semantic overlap in targeted evaluation. We view the collection of high-quality reference summaries as a valuable, orthogonal direction to this work, and we plan to explore ACU weighting in future work.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. We are grateful to Arman Cohan for insightful discussions and suggestions, Daniel Deutsch for the initial discussions, Richard Yuanzhe Pang for sharing system outputs, and Philippe Laban for valuable comments.

References

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813, Portorož, Slovenia. European Language Resources Association (ELRA).

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021a. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021b. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.

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.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2022a. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9(0):391–409.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022b. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1479–1491, Online. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *ArXiv*, abs/2302.04166.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. 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, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *ArXiv preprint*, abs/2202.06935.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

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.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization.

Pengcheng He, Baolin Peng, Liyang Lu, Song Wang, Jie Mei, Yang Liu, Ruochen Xu, Hany Hassan Awadalla, Yu Shi, Chenguang Zhu, et al. 2022. Z-code++: A pre-trained language model optimized for abstractive summarization. *ArXiv preprint*, abs/2208.09770.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Dragomir Radev, Yejin Choi, and Noah A Smith. 2022a. Beam decoding with controlled patience. *ArXiv preprint*, abs/2204.05424.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. 2022b. Bidimensional leaderboards: Generate and evaluate language hand in hand. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3540–3557, Seattle, United States. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 20d. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *ArXiv preprint*, abs/d.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Chin-Yew Lin. 2004a. Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough? In *NTCIR*.

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

Feifan Liu and Yang Liu. 2008. Correlation between ROUGE and human evaluation of extractive meeting

summaries. In *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *ArXiv*, abs/2303.16634.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Ye Ma, Zixun Lan, Lu Zong, and Kaizhu Huang. 2021. Global-aware beam search for neural abstractive summarization. In *Advances in Neural Information Processing Systems*, volume 34, pages 16545–16557. Curran Associates, Inc.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, 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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 20d. Planning with Learned Entity Prompts for Abstractive Summarization. *ArXiv preprint*, abs/d.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Karolina Owczarzak, Peter A. Rankel, Hoa Trang Dang, and John M. Conroy. 2012. Assessing the effect of inconsistent assessors on summarization evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 359–362, Jeju Island, Korea. Association for Computational Linguistics.

Richard Yuanzhe Pang and He He. 2021. Text generation by learning from demonstrations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. 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.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaiqiang Song, Bingqing Wang, Zhe Feng, and Fei Liu. 2021. A new approach to overgenerating and scoring abstractive summaries. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1392–1404, Online. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2021. How to evaluate a summarizer: Study design and statistical analysis for manual linguistic quality evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1861–1875, Online. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2022. Evaluating the factual consistency of large language models through summarization. *ArXiv preprint*, abs/2211.08412.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. 2022a. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors.

Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022b. Investigating crowdsourcing protocols for evaluating the factual consistency of summaries. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5680–5692, Seattle, United States. Association for Computational Linguistics.

Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436.

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.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Johnny Wei and Robin Jia. 2021. The statistical advantage of automatic NLG metrics at the system level. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6840–6854, Online. Association for Computational Linguistics.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5108–5122, Online. 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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020b. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020d. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

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.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation.

A Benchmark Data Collection

A.1 Detailed Settings

We discuss the detailed settings of ACU collection in §3.2. To ensure the consistency of written ACUs among different annotators, we require each annotator to be familiar with the annotation protocol and proofread each other's annotations to resolve any differences in initial annotations. After establishing a consistent understanding of the task, we have each reference summary annotated by one annotator. We note that there are multiple valid ways of writing the same atomic fact. In preliminary protocol analysis, we had multiple annotators write ACUs for the same reference summaries and did not find large differences in downstream inter-annotator agreement for ACU matching. The average time to write ACUs of one summary ranges from 2 to 5 minutes, and the overall annotation time for ACU writing is around 150 hours.

We use the following qualifications, in addition to a qualification test, to recruit MTurk workers with good track records: HIT approval rate greater than or equal to 98%, number of HITs approved greater than or equal to 10000, and located in either the United Kingdom or the United States. Workers were compensated between $0.15 and $0.55 per summary-level ACU HITs, with HITs bucketed according to the number of ACUs to be matched. For protocol comparison HITs, workers were compensated between $1 and $3. All HITs were carefully calibrated to equal a $12/hour pay rate.

The datasets we used for the collection are CNNDM, XSum and SamSum. The data release licenses are the Apache License for CNNDM and XSum, and CC BY-NC-ND 4.0 for SamSum. Our collected benchmark will be released under the 3-Clause BSD license.

A.2 Summarization Models

We list the summarization models for ACUs annotations on CNNDM, XSum, and SamSum in §3.3.

CNNDM Systems:

BART (Lewis et al., 2020b) introduce a denoising autoencoder for pretraining sequence to sequence tasks which is applicable to both natural language understanding and generation tasks.

Pegasus (Zhang et al., 2020b) introduce a model pretrained with a novel objective function designed for summarization by which important sentences are removed from an input document and then generated from the remaining sentences.

MatchSum (Zhong et al., 2020) propose a summary-level extractive system using semantic match between the extracted summary and the source document.

CTRLSum (He et al., 2020) introduce a method for controllable summarization based on keyword or descriptive prompt control tokens.

CLIFF (Cao and Wang, 2021) propose to use contrastive learning to improve factual consistency. We use the CLIFF output that uses an underlying BART model.

GOLD (Pang and He, 2021) frames text generation as an offline reinforcement learning problem, using importance weighting and assigning weights to examples that receive a higher probability from the generation model.

GSum (Dou et al., 2021) is a framework for incorporating forms of summarization guidance.

SimCLS (Liu and Liu, 2021) is a two-stage summarization model where candidates from BART are reranking by a RoBERTa (Liu et al., 2019) scoring model trained using contrastive learning.

FROST (Narayan et al., 20d) propose to do content planning in both pretraining and finetuning summarization models with plans in the form of entity chains.

GLOBAL (Ma et al., 2021) propose a variation of beam search that takes into account the global attention distribution.

BRIO (Liu et al., 2022) proposes to train a summarization model both as a token-level generator and an evaluator of sequence candidates through contrastive reranking.

BRIO-Ext (Liu et al., 2022) uses BRIO's reranker on candidate extractive summaries from MatchSum.

The following models were included in protocol-comparison annotations.

T0 (Sanh et al., 2022) introduces a prompt-based model that is fine-tuned on multiple tasks, including summarization.

GPT-3 (Brown et al., 2020) is the davinci-002 model trained on human demonstrations and model outputs highly rated by humans.[13]

XSum Systems For XSum we reuse several of the above models with their XSum-trained checkpoints as well as several variations from the above paper due to the scarcity of widely-available, easily-reproducible XSum outputs.

BART (Lewis et al., 2020b)

Pegasus (Zhang et al., 2020b)

CLIFF (Cao and Wang, 2021)

CLIFF-Pegasus (Cao and Wang, 2021) is the CLIFF algorithm applied with Pegasus as the underlying model.

FROST (Narayan et al., 20d)

BRIO (Liu et al., 2022)

BRIO-ranking (Liu et al., 2022) is the paper's reranking model.

BART-beam-patience (Kasai et al., 2022a)

SamSum Systems

We use system outputs from Gao and Wan (2022).

BART (Lewis et al., 2020b)

Pegasus (Zhang et al., 2020b)

[13] https://beta.openai.com/docs/model-index-for-researchers

System	ACU	*n*ACU	Len	R1F
PATIENCE (Kasai et al., 2022a)	27.11	26.59	25.00	45.07
CLIFF$_P$ (Cao and Wang, 2021)	25.13	24.94	21.29	46.20
BRIO-Mul (Liu et al., 2022)	26.34	26.15	21.15	48.73
BART (Lewis et al., 2020a)	23.99	23.78	20.98	45.56
PEGASUS (Zhang et al., 2020a)	24.83	24.67	20.21	46.84
CLIFF$_B$ (Cao and Wang, 2021)	22.09	21.93	20.17	44.52
FROST (Narayan et al., 20d)	27.93	27.77	19.86	47.83
BRIO-Ctr (Liu et al., 2022)	26.42	26.29	19.65	48.06

Table 9: Summarization system analysis on the XSum dataset. **ACU** is the ACU score (Eq. 1), ***n*ACU** is the normalized ACU score (Eq. 2), **Len** is the average summary length, and **R1F** is the ROUGE1 F1 score. **ACU** and ***n*ACU** are calculated on the 500 annotated examples (the value is multiplied by 100) while **Len** and **R1F** are calculated on the entire test set. The systems are sorted by **Len**. CLIFF$_P$ is based on PEGASUS, while CLIFF$_B$ is based on BART.

System	ACU	*n*ACU	Len	R1F
MV-BART (Chen and Yang, 2020)	47.65	33.01	23.57	53.80
Ctrl-DiaSumm (Liu and Chen, 2021)	49.05	37.20	22.97	56.33
PLM-BART (Feng et al., 2021)	43.74	32.61	21.43	53.46
S-BART (Chen and Yang, 2020)	34.57	25.95	21.01	50.36
CODS (Wu et al., 2021)	38.40	33.41	20.11	52.48
BART (Lewis et al., 2020b)	42.85	34.05	19.62	52.30
UniLM (Dong et al., 2019)	32.74	26.10	18.84	49.20
PEGASUS (Zhang et al., 2020b)	37.02	31.99	17.32	50.87

Table 10: Summarization system analysis on the SamSum dataset. **ACU** is the ACU score (Eq. 1), ***n*ACU** is the normalized ACU score (Eq. 2), **Len** is the average Summary length, and **R1F** is the ROUGE1 F1 score. **ACU** and ***n*ACU** are calculated on the 500 annotated examples (the value is multiplied by 100) while **Len** and **R1F** are calculated on the entire test set. The systems are sorted by **Len**.

UniLM (Dong et al., 2019) is a model pretrained on unidirection, bidirection, and sequence-to-sequence language modeling tasks.

Ctrl-DiaSumm (Liu and Chen, 2021) propose controlled generation using named entity plans.

PLM-BART (Feng et al., 2021) use DialogGPT (Zhang et al., 2020d) to annotate input dialogues before finetuning.

CODS (Wu et al., 2021) propose a two-stage generation model that first generates a sketch that is then used as a signal to the second-stage summarizer.

MV-BART (Chen and Yang, 2020) propose a multi-view encoder and a decoder that attends to these conversation views.

S-BART (Chen and Yang, 2020) encodes utterances as well as action and discourse graphs and introduces a decoder that attends to these different levels of granularity.

A.3 ACU Scores of Summarization Models

We report the ACU scores of the summarization systems we annotated on the XSum and SamSum datasets (§4.2) in Tab. 9 and Tab. 10, respectively. The results on CNNDM can be found in Tab. 3. For the normalized ACU score (Eq. 2), we set the normalization strength α to 2, 5, 0.5, on CNNDM, XSum, SamSum, respectively, by a grid search for de-correlating the summary length and the normalized score at the summary level.

B Power Analysis

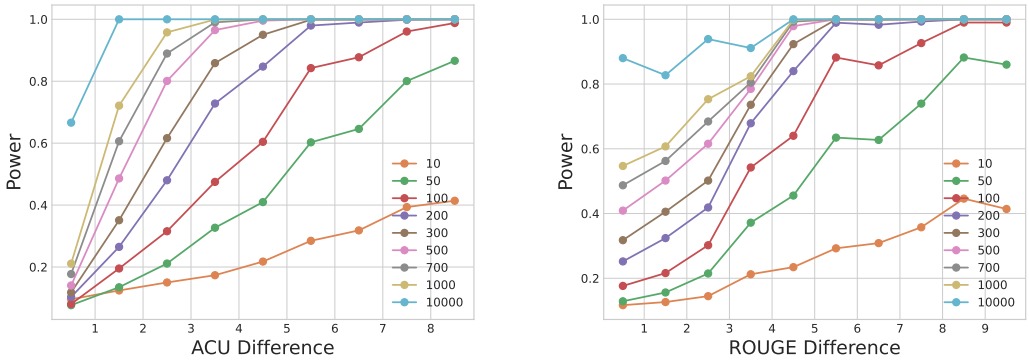
B.1 Detailed Settings

We describe the algorithm for the power analysis in §4.1 in Alg.1. While prior work (Card et al., 2020; Wei and Jia, 2021) uses parametric methods to estimate statistical power, we conduct the power analysis with the bootstrapping test (Tibshirani and Efron, 1993) as recent work (Deutsch et al., 2021b) has shown that the assumptions of the parametric methods do not always hold for human evaluation of text summarization. The process involves (1) iteratively sampling a set of examples with a certain sample size from an existing dataset, (2) running the significance test on the sampled set, and (3) estimating the power by averaging across the trials.

The essence of the test is to have a series of *simulated* datasets sampled from the existing dataset and run the significance test on the sampled sets. Here the existing dataset X consists of human-annotated scores of system outputs. We use paired bootstrapping for the significance test. The power analysis is conducted over all the system pairs.

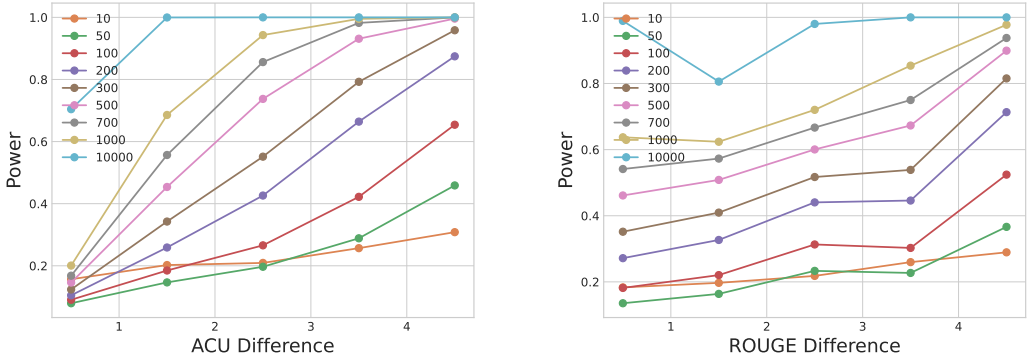
B.2 Powers of ACU Annotations

Fig.4, Fig.5, and Fig.6 show the power analysis results on CNNDM, XSum and SamSum respectively in §4.1, where the system pairs are grouped by their performance difference in either ACU or ROUGE1 recall scores. Similar to our findings on CNNDM in §4.1, we observe that increasing the sample size can effectively raise the statistical power.



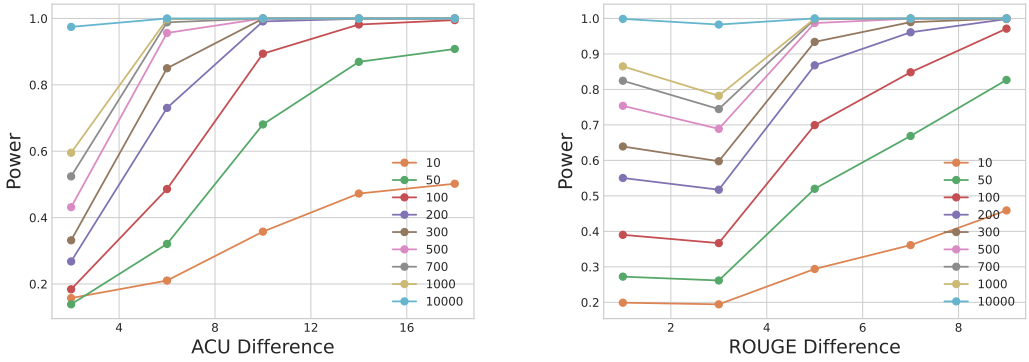
(a) (b)

Figure 4: Power analysis of human evaluation for system comparison on the annotated CNNDM test examples. Different lines represent results with different sample sizes. The system pairs are grouped by performance differences in ACU scores in Fig.4a, and by ROUGE1 recall scores in Fig.4b.



(a) (b)

Figure 5: Power analysis of human evaluation for system comparison on the annotated XSum test examples. Different lines represent results with different sample sizes. The system pairs are grouped by performance differences in ACU scores in Fig.5a, and by ROUGE1 recall scores in Fig.5b.



(a) (b)

Figure 6: Power analysis of human evaluation for system comparison on the annotated SamSum test examples. Different lines represent results with different sample sizes. The system pairs are grouped by performance differences in ACU scores in Fig.6a, and by ROUGE1 recall scores in Fig.6b.

Algorithm 1 Power Analysis

Input: n (Sample Size)
Input: X (Existing Dataset)
Input: m (Trial Number)
Output: p (Statistical Power)

```
p ← 0
for i = 0 to m do
    X̂ ← Sampling n examples from X with replacement
    p̃ ← Running the significance test on X̂
    if p̃ < 0.05 then
        p ← p + 1
    end if
end for
p ← p/m return p
```

C Calculating Correlations

We use correlations to analyze the inherent similarity between different human evaluation protocols, and the performance of automatic metrics, which is evaluated based on the correlations between the metric-calculated summary scores and the human-annotated summary scores. Specifically, given m system outputs on each of the n data samples and two different evaluation methods (e.g., human evaluation and an automatic metric) resulting in two n-row, m-column score matrices X and Y, the summary-level correlation is an average of sample-wise correlations:

$$r_{sum}(X, Y) = \frac{\sum_i \mathcal{C}(X_i, Y_i)}{n}, \quad (3)$$

where X_i, Y_i are the evaluation results on the i-th data sample and $\mathcal{C}$ is a function calculating a correlation coefficient (e.g., the Pearson correlation coefficient). In contrast, the system-level correlation is calculated on the aggregated system scores:

$$r_{sys}(X, Y) = \mathcal{C}(\bar{X}, \bar{Y}), \quad (4)$$

where $\bar{X}$ and $\bar{Y}$ contain m entries which are the system scores from the two evaluation methods averaged across n data samples, e.g., $\bar{X}_0 = \sum_i X_{i,0}/n$.

D Protocol Comparison

D.1 Data Collection Details

The 100 examples chosen for annotation in §5 are a subset of the CNNDM ACU test set, and as here we aim to analyze trends among protocols as opposed to observing statistically significant differences among systems, we believe 100 examples suffice for this collection.

Protocol	w/ Doc	w/ Ref	Fine-grained
Prior	✗	✗	✗
Ref-free	✓	✗	✗
Ref-based	✗	✓	✗
ACU	✗	✓	✓

Table 11: Human evaluation protocol comparison. We categorize the different protocols based on if they (1) require the input document (**w/ Doc**), (2) require the reference summary (**w/ Ref**), and (3) are **fine-grained**.

We summarize and compare different protocols in Tab. 11. We provide the following instructions to annotators for non-ACU annotations. We will release the full interface and instructions.

Prior: We ask the annotator to imagine each of the candidate summaries to be evaluated as a summary of a longer news article and answer the following question: how good do you think this summary is?

Ref-free: The rating measures how well the summary captures the key points of the news article. Consider whether all and only the important aspects are contained in the summary.

Ref-based: The rating measures how similar two summaries are. The similarity depends on if the summaries contain similar information, not if they use the same words.

D.2 Results Analysis

We present the result analysis of §5.2 here.

Summary Level Correlation We show the summary-level Pearson's Correlation Coefficients among different protocols in Tab. 12.

Power Analysis The power analysis results on the *Prior*, *Ref-free*, *Ref-based*, and ACU protocols are shown in Fig. 7.

Case Study We show a case study in Tab. 13 comparing the summaries generated by BRIO and GPT-3. GPT-3 scores higher on *Prior* and *Ref-free* (3.33/3.33 for BRIO and 3.66/4.00 for GPT-3). However, the BRIO summary scores 0.77 on unnormalized ACU annotations while GPT-3 scores 0.33. Also, *Ref-based* annotations favor BRIO over GPT-3 (3.66 vs. 3.33).

E Metric Analysis

E.1 Metrics

We provide additional metric details as well as results for other metrics in §6. Note that for ROUGE,

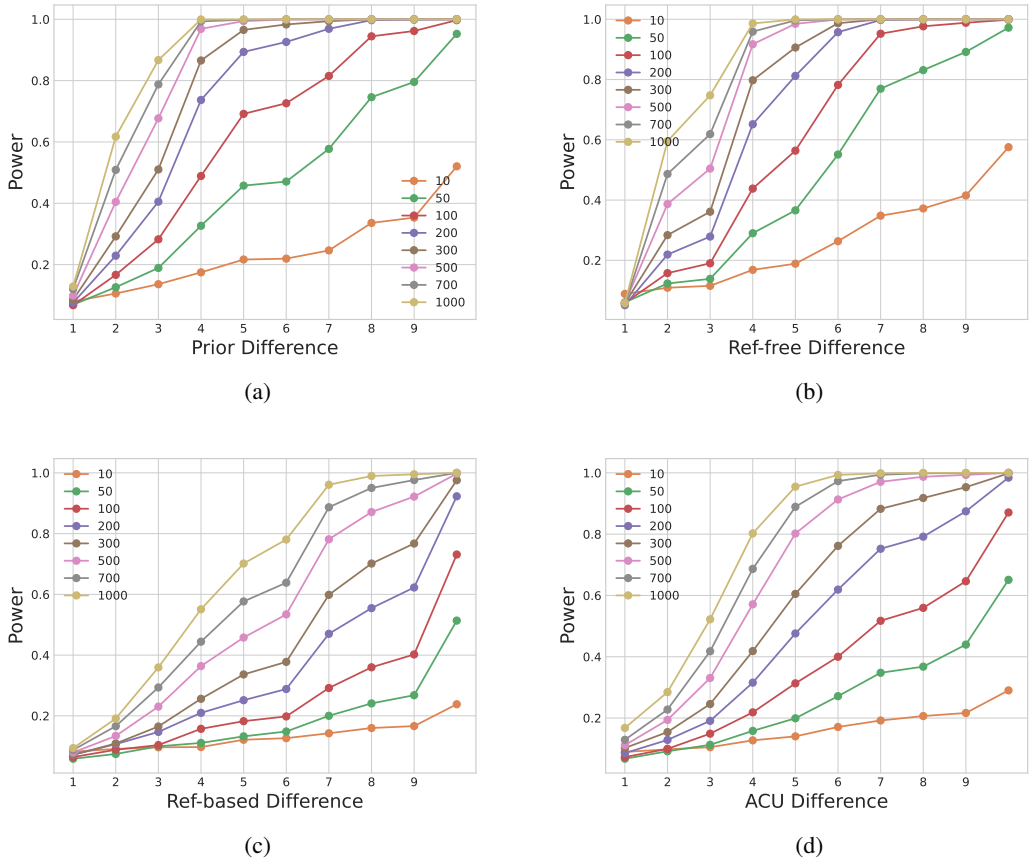


Figure 7: Power analysis of human evaluation for system comparison under *different evaluation protocols* on the annotated CNNDM test examples. Different lines represent results with different sample sizes. The system pairs are grouped into 10 buckets with similar sizes based on their performance difference under human evaluation. Fig.7a corresponds to the *Prior* protocol, Fig.7b the *Ref-free* protocol, Fig.7c the *Ref-based* protocol, and Fig.7d the ACU protocol with normalized ACU scores.

	Prior	**Ref-free**	**Ref-based**	***n*ACU**
Prior	-	**0.526**	0.056	0.082
Ref-free	0.526	-	**0.070**	0.075
Ref-based	0.056	0.070	-	0.355
*n*ACU	0.082	0.075	0.355	-
Len.	0.431	0.545	-0.107	-0.007

Table 12: *Summary-level* Pearson correlations between different protocols on the fine-tuned models. ***n*ACU** is the normalized ACU score. **Len.** is the Summary length.

we use the Python implementation. [14]

BLEU (Papineni et al., 2002) is a corpus-level precision-focused metric that calculates n-gram overlap and includes a brevity penalty.

CIDEr (Vedantam et al., 2015) computes {1-4}-gram co-occurrences, down-weighting common n-grams and calculating cosine similarity between the n-grams of the candidate and reference texts.

Statistics (Grusky et al., 2018) reports summary statistics such as the length, novel and repeated n-grams in the summary, the compression ratio between the summary and article, and measures of the level of extraction. Coverage is the percentage of words that are part of an extractive fragment and density is the average length of the extractive fragment each summary word belongs to.

MoverScore (Zhao et al., 2019) measures semantic distance with Word Mover's Distance (Kusner et al., 2015) on pooled BERT n-gram embeddings.

SUPERT (Gao et al., 2020) measures the semantic similarity of summaries with pseudo-reference summaries created by extracting salient sentences from the source documents.

BLANC (Vasilyev et al., 2020) measures the performance gains of a pre-trained language model on language understanding tasks on the input document when given access to a document summary.

[14] https://pypi.org/project/ROUGE-score/

(a) Reference Summary: Chelsea weren't awarded a penalty for David Ospina's clash with Oscar. Arsenal goalkeeper clattered Oscar inside the box. Brazilian was taken off at half-time, with Didier Drogba replacing him.

(b) System Summary (BRIO, (Liu et al., 2022)): Oscar collided with Arsenal goalkeeper David Ospina in the 16th minute of the London derby . The Brazilian was substituted at half-time and Jose Mourinho said he suffered 'possible concussion' . Oscar was knocked back by the goalkeeper but Michael Oliver didn't award Chelsea a penalty .

(c) System Summary (GPT-3, (Brown et al., 2020)): Oscar was forced to leave the match against Arsenal after sustaining a possible concussion from a collision with the opposing goalkeeper. The referee did not award Chelsea a penalty, despite the collision appearing to warrant one. Sky Sports pundits agreed that the collision should have been penalized, with some suggesting it could have even warranted a red card.

(d) ACUs with corresponding evaluations:

- Chelsea weren't awarded a penalty. ✓ ✓
- David Ospina clashed with Oscar. ✓ ✓
- David Ospina clattered Oscar. ✓ ✓
- David Ospina plays for Arsenal. ✓ ✗
- David Ospina is a goalkeeper. ✓ ✗
- The clash occurred inside the box. ✗ ✗
- Oscar is Brazilian. ✓ ✗
- Oscar was taken off at half time. ✓ ✗
- Didier Drogba replaced Oscar. ✗ ✗

Table 13: Example of a reference summary, system summaries and corresponding ACU annotations on CNNDM. The presence or absence of the ACUs for BRIO (in blue) and GPT-3 (in green) are marked by (✓) and (✗).

QAEval (Deutsch et al., 2021a) reports both an F1 and exact match (em) score. We do not report the learned answer overlap metric.

SummaQA (Scialom et al., 2019) reports an F1 score and model confidence. We plan to report QuestEval (Scialom et al., 2021) in a future version.

Lite^{3}Pyramid includes four variations of the metric depending on the entailment model (two vs three-class entailment model) and how the output is used (as a probability vs a 0/1 label).

CTC (Deng et al., 2021) proposes metrics for Compression, transduction, and creation tasks as variations of textual alignment. Relevance is scored as the average bi-directional alignment between generated and reference summaries.

SimCSE (Gao et al., 2021) apply contrastive learning to learn improved sentence representations, which can then be used to compare generated and reference summary similarity.

UniEval (Zhong et al., 2022) frames text evaluation as the answer to yes or no questions, in our case whether the summary is relevant or not, and constructs pseudo-data to fine-tune language models for this setting.

E.2 Metrics based on Large Language Models

In §6.1 we evaluate two different LLM-based automatic evaluation methods.

GPTScore (Fu et al., 2023) formulates the text evaluation as the text-filling task and takes the token probability predicted by the LLMs as the quality score. We use the following prompt for calculating the recall score of the system outputs:

> Answer the question based on the following reference summary and candidate summary.
>
> Question: Can all of the information in the reference summary be found in the candidate summary? (a). Yes. (b). No.
>
> Reference Summary: {{Reference}}
>
> Candidate Summary: {{Candidate}}
>
> Answer: Yes

The LLM-predicted probability of the last token, "Yes", is used as the recall score. We use the OpenAI's `text-davinci-003` as the LLM.

G-Eval (Liu et al., 2023) introduces a similar task as **GPTscore**, but has the LLM to predict a numerical score directly instead of using the LLM-predicted probability. We use the following prompt for the task:

> You will receive a reference summary and a candidate summary. Your task is to compare these two summaries and assess the extent to which the candidate summary covers the information presented in the reference summary.
>
> Please indicate your agreement with the following statement: "All of the information in the reference summary can be found in the candidate summary."
>
> Use the following 5-point scale when determining your response:
>
> 1. Strongly Disagree

2. Disagree

3. Neither Agree nor Disagree

4. Agree

5. Strongly Agree

Input:

Reference Summary:

{{Reference}}

Candidate Summary:

{{Candidate}}

Evaluation Form (scores ONLY):

- Agreement (1-5):

We note that we set the sampling temperature to 0 to ensure more deterministic behavior for G-Eval-3.5 and G-Eval-4. We also experiment with a sampling strategy with GPT-3.5 (G-Eval-3.5-S), where we sample 5 outputs with a temperate 1 and take the average score as the final prediction.

E.3 Metric Correlation with ACU Scores

We collect in total 50 different automatic metrics (including different variations of the same metric), and evaluate their performance using our collected ACU benchmark on CNNDM, XSum and SamSum datasets with three different correlation coefficients (§6.1). Tab. 15 reports the ***system-level*** correlation with the *un-normalized* ACU score (Eq. 1). Tab. 16 reports the ***summary-level*** correlation with the *un-normalized* ACU score (Eq. 1). Tab. 17 reports the ***system-level*** correlation with the *normalized* ACU score (Eq. 2). Tab. 18 reports the ***summary-level*** correlation with the *normalized* ACU score (Eq. 2).

E.4 System Pairs for Fine-grained Metric Evaluation

For metric elevation in §6.1, we provide the system pairs in the six different buckets grouped by their performance differences below.

Bucket 1: CLIFF V.S. FROST, CTRLSUM V.S. GSUM, BART V.S. CLIFF, GOLD V.S. FROST, BART V.S. FROST, CLIFF V.S. GOLD, BRIO V.S. GSUM, GOLD V.S. PEGASUS, BRIO V.S. CTRLSUM, BART V.S. GOLD, BRIO-EXT V.S. MATCHSUM.

Bucket 2: FROST V.S. PEGASUS, CLIFF V.S. PEGASUS, PEGASUS V.S. GLOB, BRIO-EXT V.S. SIMCLS, BART V.S. PEGASUS, BRIO V.S. MATCHSUM, BART V.S. SIMCLS, GOLD V.S. GLOB, CLIFF V.S. SIMCLS, MATCHSUM V.S. GSUM, SIMCLS V.S. FROST.

Bucket 3: MATCHSUM V.S. SIMCLS, FROST V.S. GLOB, MATCHSUM V.S. CTRLSUM, CLIFF V.S. GLOB, BRIO V.S. BRIO-EXT, GOLD V.S. SIMCLS, BART V.S. GLOB, BRIO-EXT V.S. GSUM, BRIO-EXT V.S. CTRLSUM, BART V.S. BRIO-EXT, SIMCLS V.S. PEGASUS.

Bucket 4: CLIFF V.S. BRIO-EXT, BRIO-EXT V.S. FROST, BRIO V.S. SIMCLS, GOLD V.S. BRIO-EXT, BART V.S. MATCHSUM, CLIFF V.S. MATCHSUM, SIMCLS V.S. GSUM, MATCHSUM V.S. FROST, SIMCLS V.S. GLOB, SIMCLS V.S. CTRLSUM, BRIO-EXT V.S. PEGASUS.

Bucket 5: GOLD V.S. MATCHSUM, MATCHSUM V.S. PEGASUS, BART V.S. BRIO, BRIO-EXT V.S. GLOB, BRIO V.S. CLIFF, BRIO V.S. FROST, BART V.S. GSUM, BART V.S. CTRLSUM, BRIO V.S. GOLD, CLIFF V.S. GSUM, FROST V.S. GSUM.

Bucket 6: CLIFF V.S. CTRLSUM, MATCHSUM V.S. GLOB, CTRLSUM V.S. FROST, GOLD V.S. GSUM, BRIO V.S. PEGASUS, GOLD V.S. CTRLSUM, PEGASUS V.S. GSUM, CTRLSUM V.S. PEGASUS, BRIO V.S. GLOB, GLOB V.S. GSUM, CTRLSUM V.S. GLOB.

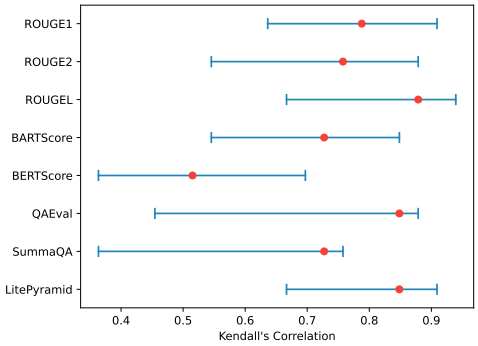


Figure 8: Confidence intervals of the system-level Kendall's correlation coefficients between automatic metrics and ACU scores.

E.5 Confidence Interval

We select several automatic metrics and calculate the confidence intervals of their system-level correlations with the ACU scores (§6.2). The results are in Fig. 8. Similar to Deutsch et al. (2021b), we found that the confidence intervals are large. However, having a larger sample size can effectively reduce the confidence interval. Specifically, we use re-sampling to generate a series of synthetic sample

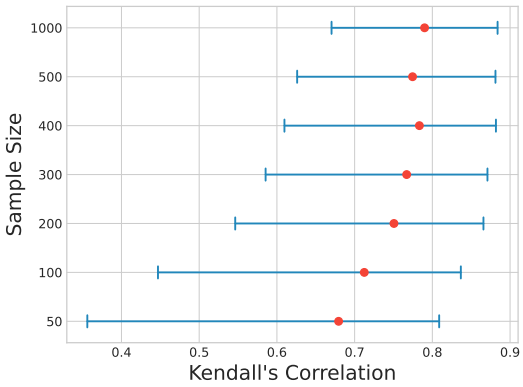


Figure 9: Confidence intervals of the system-level Kendall's correlation coefficients between ROUGE1 recall scores and ACU scores under different sample sizes.

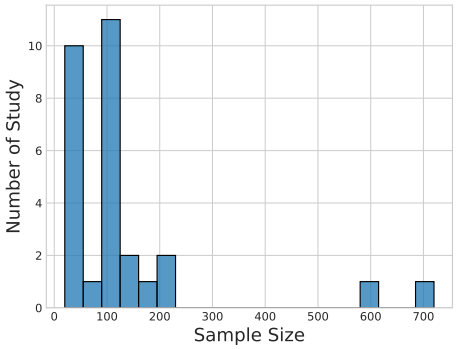


Figure 10: Sample size of the conducted human evaluation study in recent text summarization research based on our survey (Appendix F).

sets with several different sizes and calculate the confidence interval by averaging over the sampled sets with the same size. As shown in Fig. 9, larger sample sizes lead to more stable results.

E.6 Power Analysis of Metric Comparison

We use Alg.1 to conduct a power analysis of metric comparison based on their Kendall's correlations with ACU scores (§6.2). We choose 20 metrics for comparison, resulting in 190 metric pairs in total, which are (1) BARTScore-r-parabank, (2) BERTScore-r-deberta, (3) BERTScore-r-roberta, (4) BLANC, (5) CHRF, (6) CTC, (7) Meteor, (8) Lite^{2}Pyramid-p2c, (9) QAEval-em, (10) QAEval-f1, (11) ROUGE1, (12) ROUGE1r, (13) ROUGE2, (14) ROUGE2r, (15) ROUGEL, (16) ROUGELr, (17) SimCSE, (18) SummaQA, (19) SummaQA-prob, (20) SUPERT. We note that we use the *permutation test* instead of the *paired bootstrapping test* to calculate the statistical significance for metric comparison, since Deutsch et al. (2021b) found that the permutation test works better for detecting significant results in metric comparison.

E.7 Metric Correlation with Different Human Evaluation Protocols

We present the correlations between automatic metrics and different human evaluation protocols in Tab. 19 as discussed in §6.2.

F Human Evaluation Practices in Recent Text Summarization Research

We provide a brief survey for the human evaluation practices of 55 selected papers on text summarization published at NAACL[15], ACL[16], and EMNLP[17] from 2022. We follow the design of a similar study in Gehrmann et al. (2022) as described below. The results are shown in Tab. 14.

Performed Human Evaluation: Report "yes", if a human evaluation of any kind is done. We report that 71% of analyzed papers did human evaluation.

Significance Test: Report "yes", if a significance test is done on the human annotation results. Of the 39 papers that conducted human evaluation, a total of 27 papers reported the result of a significance test (68%), which is much higher compared to the 25% reported in the previous survey (Gehrmann et al., 2022).

Power Analysis: Report "yes", if a power analysis of any kind is mentioned. Of the 39 papers that conducted human evaluation, none of the papers did power analysis, the same as the result provided in the previous survey of Gehrmann et al. (2022).

Inter-annotator Agreement: Report "yes", if any kind of agreement test is conducted to evaluate the quality of human annotation themselves. Overall, we report a total of 12 papers (28%) that did agreement tests and documented specific agreement values. 9 out of the 12 papers recorded the specific agreement test, with Krippendorff's alpha as the most commonly used measurement.

Participants (crowd-worker, expert, etc.) Report "yes", if at least the number of human evaluators, document sample size, annotators per document, or their demographics is mentioned. We show the sample size of the conducted human evaluation

[15] https://aclanthology.org/events/naacl-2022/
[16] https://aclanthology.org/events/acl-2022/
[17] https://preview.aclanthology.org/emnlp-22-ingestion/volumes/2022.emnlp-main/

Best Practice & Implementation	Yes	No	Percentage (%)
Performed Human Evaluation	39	16	71
Produce Robust Human Evaluation Result			**(out of 39)**
Performed Significance Test	27	12	69
Performed Power Analysis	0	39	0
Reported Inter-annotator Agreement	12	27	28
Document Specific Agreement Test	9	3	75
Document Agreement Value	12	0	100
Documentation of the study setup (questionnaire, sample answers, platform, etc.)	39	0	100
Participants (crowd-worker, expert, etc.)	37	2	97
Document Sample Size	29	10	74
Document Participant Number	35	4	90
Document Participant Demographics	24	15	62
Released Human Evaluation Data	4	35	10

Table 14: Survey of human evaluation practices in recent text summarization research.

study in Fig. 10, and note that around 93% of them are less or equal to 200.

Released Human Evaluation Data: Report "yes", if the authors release the human evaluation data.

	CNNDM			XSum			SamSum		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
BARTScore-f1-cnndm	0.132	0.119	0.000	-0.198	0.095	0.071	0.907	0.952	0.857
BARTScore-f1-parabank	0.219	0.070	0.000	0.428	0.429	0.429	0.881	0.976	0.929
BARTScore-p-cnndm	-0.036	-0.189	-0.121	-0.650	-0.571	-0.429	0.777	0.571	0.429
BARTScore-p-parabank	-0.187	-0.294	-0.212	-0.387	-0.452	-0.286	0.692	0.524	0.286
BARTScore-r-cnndm	0.909	0.881	0.727	0.868	0.786	0.643	0.945	0.976	0.929
BARTScore-r-parabank	0.891	0.902	0.727	0.920	**0.881**	**0.714**	0.932	0.976	0.929
BERTScore-f1-deberta	0.062	0.119	0.000	0.543	0.429	0.429	0.849	0.786	0.643
BERTScore-f1-roberta	0.103	0.028	-0.091	0.592	0.429	0.429	0.852	0.809	0.714
BERTScore-p-deberta	-0.439	-0.510	-0.394	0.129	0.405	0.357	0.373	0.405	0.214
BERTScore-p-roberta	-0.275	-0.350	-0.273	0.172	0.333	0.286	0.374	0.381	0.214
BERTScore-r-deberta	0.649	0.552	0.424	0.878	0.762	0.571	0.951	0.952	0.857
BERTScore-r-roberta	0.750	0.713	0.515	0.920	0.786	0.571	0.939	0.952	0.857
BLANC	0.588	0.699	0.515	0.065	-0.024	0.071	0.824	0.809	0.714
BLEU	-0.184	-0.273	-0.212	0.631	0.595	0.571	0.806	0.833	0.714
CHRF	0.894	0.916	0.758	0.883	0.762	0.571	0.937	0.952	0.857
Compression	-0.711	-0.769	-0.606	-0.185	0.071	0.000	-0.699	-0.762	-0.571
Coverage	-0.013	-0.168	0.000	-0.568	-0.571	-0.429	0.719	0.809	0.643
CTC	0.516	0.692	0.485	0.399	0.405	0.214	0.964	0.976	0.929
Density	0.201	0.161	0.151	-0.415	-0.381	-0.286	0.815	0.762	0.571
Lite^{3}Pyramid-l2c	0.950	0.958	0.849	0.903	0.809	0.643	0.984	**1.000**	**1.000**
Lite^{3}Pyramid-l3c	0.952	0.951	0.849	0.914	0.809	0.643	**0.989**	**1.000**	**1.000**
Lite^{3}Pyramid-p2c	**0.953**	0.958	0.849	0.914	0.833	**0.714**	0.986	**1.000**	**1.000**
Lite^{3}Pyramid-p3c	0.950	**0.965**	**0.879**	**0.927**	0.809	0.643	0.987	**1.000**	**1.000**
Meteor	0.909	0.916	0.758	0.905	0.762	0.571	0.911	0.952	0.857
MoverScore	-0.173	-0.161	-0.121	0.674	0.571	0.500	0.820	0.833	0.714
Novel-1gram	0.072	0.224	0.000	0.608	0.452	0.357	-0.740	-0.809	-0.643
Novel-2gram	-0.013	0.063	0.000	0.578	0.619	0.429	-0.843	-0.833	-0.643
Repeated-1gram	0.723	0.643	0.333	0.172	0.095	0.143	0.399	0.286	0.214
Repeated-2gram	0.499	0.294	0.151	0.257	0.119	0.143	-0.277	-0.024	0.000
QAEval-em	0.723	0.629	0.515	0.450	0.452	0.357	0.947	0.952	0.857
QAEval-f1	0.925	0.944	0.849	0.551	0.500	0.429	0.962	0.976	0.929
ROUGE1	0.382	0.301	0.151	0.665	0.476	0.357	0.942	**1.000**	**1.000**
ROUGE1p	-0.490	-0.503	-0.394	0.195	0.405	0.357	0.164	0.191	0.071
ROUGE1r	0.947	0.937	0.788	0.767	0.857	**0.714**	0.920	0.976	0.929
ROUGE2	0.236	0.063	0.000	0.620	0.500	0.429	0.889	0.905	0.786
ROUGE2p	-0.412	-0.455	-0.333	0.323	0.429	0.429	0.535	0.571	0.357
ROUGE2r	0.923	0.909	0.758	0.888	0.786	0.643	0.977	**1.000**	**1.000**
ROUGEL	0.206	0.091	-0.030	0.572	0.429	0.429	0.915	0.976	0.929
ROUGELp	-0.483	-0.566	-0.424	0.181	0.357	0.286	0.317	0.381	0.214
ROUGELr	0.944	0.958	**0.879**	0.836	0.809	0.643	0.932	0.976	0.929
SimCSE	0.816	0.853	0.636	0.865	0.809	0.571	0.924	**1.000**	**1.000**
SummaQA	0.810	0.853	0.697	-0.199	-0.119	0.000	0.717	0.595	0.429
SummaQA-prob	0.749	0.860	0.727	0.308	0.214	0.143	0.817	0.738	0.643
Summary-length	0.780	0.818	0.667	0.226	-0.071	0.000	0.699	0.762	0.571
SUPERT	0.406	0.552	0.424	0.004	-0.095	-0.071	0.673	0.691	0.429
UniEval-coherence	-0.325	0.126	0.000	0.095	0.095	0.071	0.702	0.786	0.643
UniEval-consistency	0.001	-0.168	-0.061	0.056	0.071	0.071	0.344	0.238	0.071
UniEval-fluency	0.249	0.420	0.273	-0.703	-0.643	-0.500	0.405	0.381	0.286
UniEval-overall	-0.201	0.028	0.030	-0.062	0.048	0.071	0.739	0.571	0.500
UniEval-relevance	-0.148	0.119	0.091	0.017	0.333	0.214	0.742	0.643	0.500

Table 15: The ***system-level*** Pearson's r, Spearman's ρ, and Kendall's τ correlation coefficients between the automatic metric scores and ***un-normalized*** ACU scores of system outputs on CNNDM, XSum and SamSum datasets.

	CNNDM			XSum			SamSum		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
BARTScore-f1-cnndm	0.353	0.329	0.264	0.261	0.238	0.202	0.434	0.401	0.340
BARTScore-f1-parabank	0.417	0.386	0.311	0.309	0.279	0.239	0.430	0.396	0.340
BARTScore-p-cnndm	0.178	0.161	0.128	0.188	0.166	0.140	0.282	0.263	0.224
BARTScore-p-parabank	0.237	0.216	0.170	0.235	0.220	0.187	0.269	0.249	0.212
BARTScore-r-cnndm	0.567	0.530	0.435	0.325	0.300	0.260	0.546	0.508	0.438
BARTScore-r-parabank	0.582	0.548	0.453	**0.353**	**0.326**	0.282	0.531	0.500	0.430
BERTScore-f1-deberta	0.441	0.413	0.334	0.290	0.280	0.241	0.401	0.377	0.326
BERTScore-f1-roberta	0.432	0.397	0.320	0.305	0.285	0.244	0.415	0.388	0.335
BERTScore-p-deberta	0.255	0.239	0.191	0.209	0.211	0.180	0.208	0.204	0.173
BERTScore-p-roberta	0.218	0.200	0.160	0.223	0.221	0.190	0.212	0.209	0.178
BERTScore-r-deberta	0.544	0.516	0.424	0.327	0.305	0.262	0.507	0.476	0.409
BERTScore-r-roberta	0.571	0.542	0.448	0.348	0.320	0.277	0.516	0.481	0.417
BLANC	0.238	0.220	0.175	-0.018	-0.022	-0.020	0.167	0.156	0.136
BLEU	0.337	0.306	0.246	0.275	0.259	0.227	0.373	0.356	0.306
CHRF	0.564	0.528	0.436	0.353	0.315	0.275	0.486	0.459	0.396
Compression	-0.309	-0.296	-0.238	-0.088	-0.080	-0.071	-0.312	-0.307	-0.269
Coverage	0.012	0.005	0.003	-0.045	-0.044	-0.037	0.056	0.044	0.037
CTC	0.453	0.431	0.348	0.270	0.249	0.215	0.476	0.442	0.382
Density	0.078	0.070	0.054	-0.054	-0.052	-0.044	0.119	0.109	0.091
Lite^{3}Pyramid-l2c	0.537	0.523	0.466	0.219	0.219	0.207	0.524	0.519	0.494
Lite^{3}Pyramid-l3c	0.532	0.521	0.466	0.217	0.214	0.204	0.540	0.535	**0.509**
Lite^{3}Pyramid-p2c	0.582	0.546	0.452	0.303	0.284	0.245	0.599	0.539	0.467
Lite^{3}Pyramid-p3c	**0.584**	0.543	0.448	0.310	0.285	0.246	**0.615**	**0.549**	0.475
Meteor	0.537	0.496	0.407	0.327	0.308	0.268	0.471	0.430	0.373
MoverScore	0.388	0.364	0.292	0.320	0.296	0.252	0.398	0.375	0.320
Novel-1gram	-0.008	-0.005	-0.003	0.051	0.048	0.041	-0.070	-0.066	-0.056
Novel-2gram	-0.026	-0.035	-0.028	0.057	0.057	0.050	-0.112	-0.103	-0.087
Repeated-1gram	0.071	0.067	0.052	0.010	0.006	0.005	0.172	0.172	0.152
Repeated-2gram	0.061	0.060	0.048	0.010	0.006	0.005	0.059	0.057	0.052
QAEval-em	0.350	0.334	0.296	0.159	0.156	0.149	0.383	0.377	0.352
QAEval-f1	0.454	0.427	0.358	0.226	0.215	0.198	0.437	0.421	0.384
ROUGE1	0.457	0.430	0.348	0.302	0.292	0.253	0.416	0.398	0.345
ROUGE1p	0.190	0.175	0.140	0.227	0.224	0.194	0.113	0.119	0.103
ROUGE1r	0.579	**0.552**	**0.468**	0.328	0.322	**0.293**	0.503	0.485	0.439
ROUGE2	0.444	0.407	0.329	0.277	0.255	0.222	0.380	0.350	0.301
ROUGE2p	0.307	0.287	0.229	0.241	0.229	0.200	0.214	0.210	0.181
ROUGE2r	0.552	0.529	0.453	0.301	0.291	0.266	0.456	0.436	0.395
ROUGEL	0.430	0.399	0.321	0.266	0.249	0.215	0.395	0.372	0.323
ROUGELp	0.192	0.179	0.143	0.214	0.208	0.180	0.121	0.120	0.103
ROUGELr	0.561	0.537	0.454	0.297	0.285	0.258	0.480	0.460	0.415
SimCSE	0.461	0.429	0.346	0.308	0.290	0.248	0.450	0.420	0.360
SummaQA	0.165	0.153	0.121	0.022	0.015	0.013	0.045	0.049	0.039
SummaQA-prob	0.155	0.150	0.119	0.026	0.023	0.019	0.131	0.120	0.102
Summary-length	0.315	0.296	0.238	0.081	0.075	0.067	0.314	0.307	0.268
SUPERT	0.211	0.206	0.165	0.047	0.049	0.042	0.191	0.168	0.141
UniEval-coherence	0.098	0.127	0.100	0.017	0.011	0.012	0.186	0.197	0.167
UniEval-consistency	0.007	0.015	0.010	0.017	0.013	0.013	0.044	0.037	0.031
UniEval-fluency	-0.008	-0.022	-0.015	-0.031	-0.040	-0.034	-0.006	-0.035	-0.028
UniEval-overall	0.111	0.132	0.104	0.089	0.078	0.067	0.171	0.187	0.157
UniEval-relevance	0.129	0.154	0.121	0.180	0.181	0.152	0.201	0.235	0.197

Table 16: The ***summary-level*** Pearson's r, Spearman's ρ, and Kendall's τ correlation coefficients between the automatic metric scores and ***un-normalized*** ACU scores of system outputs on CNNDM, XSum and SamSum.

	CNNDM			XSum			SamSum		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
BARTScore-f1-cnndm	0.539	0.706	0.455	-0.148	0.095	0.071	0.920	0.786	0.643
BARTScore-f1-parabank	0.692	0.706	0.455	0.478	0.429	0.429	0.907	0.714	0.571
BARTScore-p-cnndm	0.430	0.524	0.333	-0.622	-0.571	-0.429	0.945	0.738	0.643
BARTScore-p-parabank	0.421	0.413	0.242	-0.341	-0.452	-0.286	0.915	0.691	0.500
BARTScore-r-cnndm	0.461	0.364	0.273	0.893	0.786	0.643	0.774	0.738	0.571
BARTScore-r-parabank	0.756	0.615	0.455	**0.932**	**0.881**	**0.714**	0.786	0.738	0.571
BERTScore-f1-deberta	0.643	0.783	0.576	0.593	0.429	0.429	**0.985**	0.952	0.857
BERTScore-f1-roberta	0.733	0.755	0.545	0.639	0.429	0.429	0.955	**0.976**	**0.929**
BERTScore-p-deberta	0.335	0.378	0.242	0.191	0.405	0.357	0.756	0.619	0.571
BERTScore-p-roberta	0.432	0.420	0.242	0.233	0.333	0.286	0.779	0.667	0.571
BERTScore-r-deberta	0.664	0.489	0.333	0.863	0.762	0.571	0.845	0.714	0.500
BERTScore-r-roberta	0.725	0.552	0.364	0.909	0.786	0.571	0.802	0.714	0.500
BLANC	-0.122	-0.126	-0.061	0.020	-0.024	0.071	0.506	0.452	0.357
BLEU	0.442	0.482	0.303	0.676	0.595	0.571	0.906	0.905	0.786
CHRF	0.701	0.601	0.424	0.869	0.762	0.571	0.818	0.714	0.500
Compression	-0.099	-0.077	-0.091	-0.128	0.071	0.000	-0.361	-0.357	-0.214
Coverage	-0.599	-0.797	-0.576	-0.603	-0.571	-0.429	0.606	0.381	0.286
CTC	-0.074	-0.035	-0.030	0.350	0.405	0.214	0.792	0.738	0.571
Density	-0.366	-0.685	-0.485	-0.427	-0.381	-0.286	0.574	0.286	0.214
Lite^{3}Pyramid-l2c	0.501	0.462	0.273	0.903	0.809	0.643	0.856	0.786	0.643
Lite^{3}Pyramid-l3c	0.486	0.448	0.273	0.908	0.809	0.643	0.845	0.786	0.643
Lite^{3}Pyramid-p2c	0.510	0.462	0.273	0.915	0.833	**0.714**	0.847	0.786	0.643
Lite^{3}Pyramid-p3c	0.498	0.503	0.303	0.921	0.809	0.643	0.840	0.786	0.643
Meteor	0.744	0.601	0.424	0.901	0.762	0.571	0.796	0.714	0.500
MoverScore	0.540	0.594	0.394	0.718	0.571	0.500	0.879	0.905	0.786
Novel-1gram	0.637	0.811	0.636	0.635	0.452	0.357	-0.594	-0.381	-0.286
Novel-2gram	0.593	0.769	0.576	0.612	0.619	0.429	-0.609	-0.429	-0.286
Repeated-1gram	0.357	0.224	0.182	0.119	0.095	0.143	0.102	0.000	0.000
Repeated-2gram	0.382	0.252	0.182	0.243	0.119	0.143	-0.211	-0.095	-0.071
QAEval-em	0.408	0.350	0.242	0.489	0.452	0.357	0.909	0.786	0.643
QAEval-f1	0.602	0.489	0.333	0.588	0.500	0.429	0.894	0.714	0.571
ROUGE1	**0.915**	**0.881**	**0.788**	0.704	0.476	0.357	0.909	0.786	0.643
ROUGE1p	0.272	0.329	0.182	0.257	0.405	0.357	0.548	0.524	0.429
ROUGE1r	0.516	0.413	0.273	0.737	0.857	**0.714**	0.644	0.738	0.571
ROUGE2	0.770	0.699	0.515	0.665	0.500	0.429	0.961	0.881	0.714
ROUGE2p	0.317	0.406	0.242	0.382	0.429	0.429	0.839	0.833	0.714
ROUGE2r	0.651	0.510	0.364	0.893	0.786	0.643	0.842	0.786	0.643
ROUGEL	0.811	0.790	0.606	0.620	0.429	0.429	0.897	0.857	0.714
ROUGELp	0.275	0.280	0.151	0.243	0.357	0.286	0.660	0.667	0.571
ROUGELr	0.600	0.524	0.364	0.815	0.809	0.643	0.688	0.738	0.571
SimCSE	0.801	0.685	0.545	0.876	0.809	0.571	0.847	0.786	0.643
SummaQA	0.196	0.133	0.061	-0.239	-0.119	0.000	0.334	0.071	0.071
SummaQA-prob	0.591	0.503	0.212	0.251	0.214	0.143	0.416	0.357	0.286
Summary-length	0.087	0.042	0.030	0.166	-0.071	0.000	0.329	0.357	0.214
SUPERT	-0.233	-0.329	-0.212	-0.057	-0.095	-0.071	0.293	0.238	0.071
UniEval-coherence	-0.620	-0.357	-0.273	0.062	0.095	0.071	0.481	0.333	0.286
UniEval-consistency	-0.534	-0.748	-0.515	0.023	0.071	0.071	0.606	0.214	0.000
UniEval-fluency	0.286	0.189	0.121	-0.681	-0.643	-0.500	0.667	0.500	0.357
UniEval-overall	0.178	0.126	0.121	-0.072	0.048	0.071	0.734	0.381	0.286
UniEval-relevance	0.365	0.762	0.545	0.064	0.333	0.214	0.717	0.548	0.429

Table 17: The ***system-level*** Pearson's r, Spearman's ρ, and Kendall's τ correlation coefficients between the automatic metric scores and ***normalized*** ACU scores of system outputs on CNNDM, XSum and SamSum datasets.

	CNNDM			XSum			SamSum		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
BARTScore-f1-cnndm	0.398	0.384	0.296	0.276	0.274	0.228	0.445	0.413	0.341
BARTScore-f1-parabank	0.465	0.441	0.342	0.327	0.310	0.260	0.483	0.442	0.371
BARTScore-p-cnndm	0.284	0.273	0.209	0.205	0.208	0.172	0.389	0.355	0.295
BARTScore-p-parabank	0.355	0.338	0.259	0.257	0.262	0.216	0.411	0.367	0.307
BARTScore-r-cnndm	0.485	0.435	0.334	0.329	0.303	0.257	0.439	0.411	0.345
BARTScore-r-parabank	0.507	0.462	0.357	**0.361**	0.329	0.277	0.462	0.444	0.372
BERTScore-f1-deberta	0.518	0.491	0.386	0.317	0.323	0.274	**0.517**	**0.472**	**0.401**
BERTScore-f1-roberta	0.515	0.486	0.386	0.333	**0.330**	**0.280**	0.512	0.470	0.401
BERTScore-p-deberta	0.423	0.411	0.320	0.248	0.294	0.248	0.413	0.382	0.323
BERTScore-p-roberta	0.389	0.378	0.296	0.261	0.296	0.250	0.411	0.377	0.319
BERTScore-r-deberta	0.491	0.454	0.354	0.329	0.291	0.246	0.474	0.441	0.372
BERTScore-r-roberta	0.515	0.476	0.371	0.349	0.306	0.260	0.470	0.442	0.375
BLANC	0.045	0.020	0.014	-0.031	-0.041	-0.036	0.034	0.038	0.035
BLEU	0.441	0.414	0.328	0.294	0.300	0.260	0.469	0.432	0.368
CHRF	0.527	0.479	0.379	0.351	0.301	0.256	0.438	0.412	0.351
Compression	-0.053	-0.002	0.021	-0.037	0.069	0.071	-0.037	-0.021	0.001
Coverage	-0.016	-0.020	-0.019	-0.051	-0.049	-0.040	0.055	0.045	0.037
CTC	0.349	0.317	0.237	0.274	0.231	0.194	0.414	0.385	0.326
Density	-0.031	-0.040	-0.032	-0.059	-0.067	-0.055	0.050	0.051	0.046
Lite^{3}Pyramid-l2c	0.452	0.424	0.355	0.212	0.197	0.181	0.410	0.404	0.374
Lite^{3}Pyramid-l3c	0.449	0.427	0.358	0.213	0.197	0.180	0.418	0.417	0.385
Lite^{3}Pyramid-p2c	0.482	0.420	0.321	0.294	0.259	0.216	0.462	0.419	0.353
Lite^{3}Pyramid-p3c	0.489	0.430	0.330	0.298	0.253	0.211	0.469	0.419	0.354
Meteor	0.484	0.435	0.337	0.329	0.303	0.260	0.427	0.391	0.335
MoverScore	0.509	0.483	0.380	0.341	0.329	0.275	0.513	0.459	0.386
Novel-1gram	0.017	0.020	0.018	0.054	0.045	0.037	-0.066	-0.059	-0.049
Novel-2gram	0.053	0.049	0.037	0.063	0.068	0.055	-0.060	-0.058	-0.052
Repeated-1gram	-0.029	-0.044	-0.033	-0.015	-0.030	-0.022	-0.004	0.007	0.011
Repeated-2gram	-0.012	-0.016	-0.007	0.004	-0.009	-0.007	-0.036	-0.032	-0.027
QAEval-em	0.322	0.302	0.253	0.161	0.155	0.144	0.328	0.321	0.289
QAEval-f1	0.412	0.379	0.301	0.227	0.207	0.188	0.367	0.349	0.307
ROUGE1	**0.541**	**0.510**	**0.403**	0.324	0.324	0.278	0.494	0.464	0.399
ROUGE1p	0.386	0.380	0.298	0.263	0.303	0.262	0.320	0.308	0.269
ROUGE1r	0.425	0.374	0.290	0.317	0.276	0.244	0.349	0.329	0.286
ROUGE2	0.504	0.473	0.375	0.296	0.292	0.253	0.432	0.402	0.343
ROUGE2p	0.439	0.424	0.335	0.270	0.291	0.253	0.346	0.330	0.283
ROUGE2r	0.474	0.433	0.347	0.298	0.274	0.243	0.379	0.359	0.317
ROUGEL	0.512	0.481	0.378	0.288	0.293	0.252	0.462	0.427	0.365
ROUGELp	0.380	0.375	0.294	0.249	0.287	0.248	0.313	0.293	0.254
ROUGELr	0.424	0.378	0.293	0.292	0.258	0.227	0.338	0.321	0.279
SimCSE	0.437	0.393	0.300	0.321	0.305	0.255	0.454	0.422	0.356
SummaQA	0.059	0.044	0.031	0.018	0.001	0.002	0.010	0.022	0.014
SummaQA-prob	0.076	0.069	0.050	0.019	-0.003	-0.003	0.050	0.054	0.047
Summary-length	0.040	0.002	-0.021	0.027	-0.075	-0.079	0.010	0.021	-0.001
SUPERT	0.062	0.045	0.030	0.031	0.013	0.008	0.064	0.059	0.044
UniEval-coherence	0.083	0.087	0.060	0.011	-0.018	-0.014	0.108	0.112	0.090
UniEval-consistency	0.001	-0.006	-0.009	0.008	-0.015	-0.010	0.087	0.075	0.060
UniEval-fluency	0.017	0.001	0.001	-0.031	-0.040	-0.032	0.035	0.014	0.009
UniEval-overall	0.130	0.146	0.107	0.088	0.062	0.052	0.204	0.208	0.168
UniEval-relevance	0.150	0.173	0.127	0.187	0.192	0.158	0.236	0.252	0.203

Table 18: The ***summary-level*** Pearson's r, Spearman's ρ, and Kendall's τ correlation coefficients between the automatic metric scores and ***normalized*** ACU scores of system outputs on CNNDM, XSum and SamSum datasets.

	System-level Correlation				Summary-level Correlation			
	Prior	**Ref-free**	**Ref-based**	***n*ACU**	**Prior**	**Ref-free**	**Ref-based**	***n*ACU**
BARTScore_f1_cnndm	-0.030	-0.121	0.656	0.364	0.060	0.032	0.335	0.280
BARTScore_f1_parabank	-0.091	-0.182	0.656	0.364	0.079	0.038	0.369	0.324
BARTScore_p_cnndm	-0.091	-0.182	0.595	0.242	-0.008	-0.033	0.281	0.210
BARTScore_p_parabank	-0.273	-0.364	0.534	0.242	-0.001	-0.039	0.321	0.235
BARTScore_r_cnndm	0.545	0.455	-0.076	0.212	0.159	0.162	0.290	0.281
BARTScore_r_parabank	0.394	0.364	0.199	0.485	0.175	0.192	0.292	0.323
BERTscore_f1_deberta	-0.061	-0.212	0.779	0.576	0.040	0.013	0.398	0.366
BERTscore_f1_roberta	-0.091	-0.182	0.779	0.485	0.030	-0.001	**0.399**	0.358
BERTscore_p_deberta	-0.333	-0.485	0.626	0.364	-0.077	-0.132	0.366	0.310
BERTscore_p_roberta	-0.242	-0.333	0.687	0.333	-0.101	-0.125	0.334	0.283
BERTscore_r_deberta	0.273	0.121	0.626	0.667	0.171	0.178	0.312	0.314
BERTscore_r_roberta	0.273	0.121	0.565	**0.727**	0.176	0.193	0.331	0.328
BLANC	0.545	0.576	-0.504	-0.212	0.256	0.310	-0.051	-0.006
BLEU	-0.182	-0.333	0.534	0.515	-0.025	-0.046	0.260	0.296
CHRF	0.576	0.424	0.199	0.485	0.161	0.181	0.284	0.347
Compression	-0.606	-0.758	0.382	0.091	-0.344	-0.422	0.081	0.046
Coverage	-0.030	0.121	-0.779	-0.606	0.076	0.074	-0.035	-0.019
CTC	0.455	0.545	-0.473	-0.242	0.267	0.295	0.193	0.221
Density	0.121	0.273	-0.687	-0.515	0.134	0.117	-0.082	-0.055
Lite^{3}Pyramid-l2c	0.545	0.576	-0.046	0.212	0.214	0.218	0.230	0.337
Lite^{3}Pyramid-l3c	0.545	0.636	-0.107	0.091	0.199	0.219	0.223	0.342
Lite^{3}Pyramid-p2c	0.576	0.667	-0.168	0.121	0.234	0.254	0.206	0.301
Lite^{3}Pyramid-p3c	0.576	0.667	-0.107	0.182	0.210	0.242	0.196	0.305
Meteor	0.394	0.242	0.382	0.485	0.157	0.149	0.274	0.313
MoverScore	-0.151	-0.364	**0.870**	0.545	0.009	-0.018	0.360	0.356
Novel-1gram	0.030	-0.121	0.779	0.606	-0.071	-0.066	0.037	0.018
Novel-2gram	-0.061	-0.273	**0.870**	0.636	-0.097	-0.088	0.088	0.057
Repeated-1gram	0.212	0.303	-0.321	-0.121	0.097	0.102	-0.035	-0.032
Repeated-2gram	0.000	-0.091	-0.046	0.030	0.068	0.074	-0.010	-0.007
QAEval-em	0.303	0.333	-0.107	0.030	0.087	0.100	0.131	0.227
QAEval-f1	0.485	0.515	-0.076	0.151	0.127	0.122	0.203	0.274
ROUGE1	-0.061	-0.212	0.840	0.636	0.033	-0.008	0.346	**0.377**
ROUGE1p	-0.364	-0.515	0.687	0.394	-0.158	-0.241	0.293	0.278
ROUGE1r	**0.697**	0.667	-0.107	0.182	0.272	0.322	0.211	0.248
ROUGE2	0.000	-0.151	0.595	0.636	0.004	-0.005	0.304	0.329
ROUGE2p	-0.273	-0.424	0.626	0.424	-0.095	-0.118	0.309	0.302
ROUGE2r	0.455	0.364	0.199	0.424	0.134	0.177	0.248	0.285
ROUGEL	-0.061	-0.212	0.779	0.636	0.024	0.005	0.325	0.370
ROUGELp	-0.394	-0.545	0.656	0.364	-0.148	-0.218	0.279	0.289
ROUGELr	0.606	0.576	0.046	0.333	0.234	0.296	0.211	0.263
SimCSE	0.273	0.242	0.351	0.364	0.154	0.154	0.266	0.284
SummaQA	0.636	0.606	-0.290	0.000	0.158	0.218	-0.020	0.000
SummaQA-prob	0.515	0.424	0.260	0.303	0.144	0.176	0.021	0.049
Summary-length	0.576	**0.727**	-0.473	-0.182	**0.344**	**0.422**	-0.081	-0.046
SUPERT	0.333	0.485	-0.656	-0.424	0.206	0.245	-0.026	0.029
UniEval-coherence	0.151	0.121	-0.107	-0.061	0.229	0.176	0.025	0.071
UniEval-consistency	-0.091	0.061	-0.565	-0.545	0.077	0.064	-0.045	-0.030
UniEval-fluency	0.303	0.273	0.260	0.333	0.080	0.058	0.004	0.018
UniEval-overall	0.151	0.000	0.443	0.303	0.219	0.142	0.116	0.122
UniEval-relevance	-0.061	-0.212	0.656	0.394	0.191	0.114	0.181	0.129

Table 19: The Kendall's correlation between the automatic metric and different human evaluation protocols on CNNDM dataset.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8.

☑ A2. Did you discuss any potential risks of your work?
Section 8.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3-6.

☑ B1. Did you cite the creators of artifacts you used?
Section 3.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We did not discuss this in the paper but our use case is research-based and consistent with the underlying licenses.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We use standard benchmark datasets that have been widely used so we expect the risk of their containing offensive or personal information is relatively low.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 and Appendix A.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 6.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. We didn't have experimental results with new models.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. We didn't have experimental results with new models.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. We didn't have experimental results with new models.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A.2.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3, 5.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
In the appendix and supplementary materials.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3 and Appendix A.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The data we collected contains no personal information or free-form text, therefore we consider the risk of releasing our data relatively low. The protocol for data collection follows standard MTurk procedures.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our paper has been reviewed internally for ethics concerns.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix A.