A Critical Look at Meta-evaluating Summarisation Evaluation Metrics

Xiang Dai and **Sarvnaz Karimi** and **Biaoyan Fang**
CSIRO Data61
Sydney, Australia
{dai.dai,sarvnaz.karimi,byron.fang}@csiro.au

Abstract

Effective summarisation evaluation metrics enable researchers and practitioners to compare different summarisation systems efficiently. Estimating the effectiveness of an automatic evaluation metric, termed *meta-evaluation*, is a critically important research question. In this position paper, we review recent meta-evaluation practices for summarisation evaluation metrics and find that (1) evaluation metrics are primarily meta-evaluated on datasets consisting of examples from news summarisation datasets, and (2) there has been a noticeable shift in research focus towards evaluating the faithfulness of generated summaries. We argue that the time is ripe to build more diverse benchmarks that enable the development of more robust evaluation metrics and analyze the generalization ability of existing evaluation metrics. In addition, we call for research focusing on user-centric quality dimensions that consider the generated summary's communicative goal and the role of summarisation in the workflow.

1 Introduction

The evaluation of natural language processing systems is crucial to ensure their effectiveness and reliability in real-world applications. It helps compare systems, validate whether the designed properties work as intended, understand the strengths and weaknesses of the underlying model, and often guide iterative improvements (Ribeiro et al., 2020). Although human evaluation, especially for natural language generation systems, is considered the most reliable evaluation method (Huang et al., 2020; Iskender et al., 2021; Khashabi et al., 2022), automatic evaluation metrics are more widely used due to their cost-effectiveness, ease of use, repeatability, and speed (Graham, 2015; Gehrmann et al., 2023).

In addition to assessing the performance of summarisation systems, automatic summarisation evaluation metrics are also used for other purposes during summarisation system development, such as filtering noisy datasets to improve the quality of training data (Chaudhury et al., 2022; Aharoni et al., 2023), ranking sampled candidates to output the best summary (Falke et al., 2019; Chaudhury et al., 2022), and, integrating with reinforcement learning framework as a reward function (Zhang et al., 2020b; Stiennon et al., 2020).

A critically important question is *how effective these automatic summarisation evaluation metrics are*. In other words, do the evaluation results obtained using these automatic metrics reflect the genuine quality of the summaries and summarisation systems under examination? For example, Goyal et al. (2022) conclude that existing automatic metrics cannot reliably evaluate summaries generated using instruct-tuned GPT-3 model (Ouyang et al., 2022), because they find that GTP-3 summaries receive much lower scores than state-of-the-art fine-tuned models (Liu et al., 2022) on automatic metrics while outperforming them on human evaluation using A/B testing.

Meta-evaluating summarisation evaluation metrics, especially building resources that enable assessing the automatic metrics, has become urgent and attracted significant research interest (Fabbri et al., 2020; Bhandari et al., 2020; Clark et al., 2023; Liu et al., 2023b; Laban et al., 2023). However, these resources were built and used in various ways, leading to inconsistent and confusing conclusions about the usefulness of these metrics.

In this position paper, we take a critical look at the practices of meta-evaluating summarisation evaluation metrics. Our paper is organised as follows: we first review recent meta-evaluation practices for summarisation evaluation metrics (Section 2); then, in Section 3, we discuss research trends and gaps around four critical decisions that must be made when we assess the automatic metrics, namely, *choosing data to annotate*, *defining quality dimensions*, *collecting human judgements*,

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14795–14808
November 12-16, 2024 ©2024 Association for Computational Linguistics

and *comparing automatic metrics against human judgements*. Finally, we provide recommendations in Section 5.

2 Preliminaries

The task of summarisation aims to generate a summary $\hat{y}$ given a source text x, where $\hat{y}$ encapsulates the key information in x. The summarisation evaluation metric typically takes the generated summary $\hat{y}$, optionally the source x or a (few) reference summary y, as input and produces a numeric value, which is a proxy of the overall quality or a particular dimension of quality, of $\hat{y}$.

2.1 Summarisation Evaluation Metrics

Summarisation evaluation metrics can be roughly grouped into categories based on what input data they use (e.g., source text, reference summary), what intermediate data they generate (e.g., auto-generated questions based on the source text), what underlying models they rely on (e.g., textual entailment models):

Summary-only metrics take the generated summary $\hat{y}$ as input and focus on how well the generated text can be read, e.g., free of syntactic errors or spelling errors (Mani, 2001; Goldsack et al., 2023);

Similarity-based metrics take $\hat{y}$ and one or a few reference summaries y as input and measure how similar $\hat{y}$ and y are (Lin, 2004; Zhang et al., 2020a);

Entailment-based metrics take both $\hat{y}$ and the source x as input and use entailment models to determine whether the information in $\hat{y}$ is supported by x (Laban et al., 2022; Honovich et al., 2022);

QA-based metrics use both $\hat{y}$ and x and aim to compare the factual information in $\hat{y}$ and x by eliciting answers from them for the same question (Durmus et al., 2020; Deutsch et al., 2021a);

Learning-based metrics aim to train an evaluation model, using human annotations (Aharoni et al., 2023) or weak supervision signals (Kryscinski et al., 2020; Wu et al., 2023), that directly outputs the quality score of $\hat{y}$ given x; and,

LLM-based metrics directly instruct large language models to generate the quality score of $\hat{y}$ (Tam et al., 2023; Shen et al., 2023).

2.2 Meta-evaluation of Automatic Metrics

Estimating the effectiveness and reliability of an automatic evaluation metric is a critically important research question. To distinguish it from summarisation *evaluation*, researchers usually use the term *meta-evaluation* to refer to this task, which is the focus of our position paper.

Early studies of summarisation meta-evaluation focus on assessing evaluation metrics according to their ability to distinguish between human-written and system-generate summaries (Rankel et al., 2011). However, more recently, a widely accepted belief about meta-evaluation is that an effective evaluation metric should mirror human judgements (Graham, 2015; Huang et al., 2020; Fabbri et al., 2020; Gao and Wan, 2022). This is often approximated by calculating the correlation between the evaluation results using the automatic evaluation metric X and human judgements Z across a set of summaries generated using various systems.

Assuming there are N source texts, and J summarisation systems are employed, resulting in a total of $N \times J$ output summaries. We use d_i to represent the i-th source text and s_i^j the summary generated by the j-th summarisation system on d_i. We use x_i^j to represent the score assigned to s_i^j by the evaluation metric X and z_i^j the corresponding human judgement. To measure the correlation between X and Z, a correlation function ($\mathbf{Corr}$, such as Pearson, Kendall, or Spearman) is needed.

System-level protocol aggregates the evaluation scores for a given summarisation system first via:

$$x^j = \frac{1}{N}\sum_{i=1}^{N} x_i^j, \tag{1}$$

where x^j is an approximation of the judgement of the j-th summarisation system by metric X. Similarly, the human judgement can be aggregated via:

$$z^j = \frac{1}{N}\sum_{i=1}^{N} z_i^j. \tag{2}$$

Then, the two lists of judgements, each containing J values, are taken as input to calculate the system-level correlation coefficient and the corresponding p-value:

$$r, p = \mathbf{Corr}\left(\left[x^1, \cdot, x^J\right], \left[z^1, \cdot, z^J\right]\right). \tag{3}$$

	Data	Quality dimensions	Comparison protocol
SummEval (Fabbri et al., 2020)	Model-generated and (transformed) reference summaries on news articles, such as those in CNN/DM (Nallapati et al., 2016), XSum (Narayan et al., 2018), XL-Sum (Hasan et al., 2021), and MLSum (Scialom et al., 2020)	Coherence, Faithfulness, Fluency, Relevance	Correlation
REALSumm (Bhandari et al., 2020)		Relevance	Correlation
FRANK (Pagnoni et al., 2021)		Faithfulness	Correlation
FFCI (Koto et al., 2022)		Focus, Coverage, Coherence	Correlation
FIB (Tam et al., 2023)		Factual consistency	Ranking
BUMP (Ma et al., 2023)		Faithfulness	Ranking, Classification
SEAHORSE (Clark et al., 2023)		Comprehensibility, Repetition, Grammar, Attribution, Main ideas, and Conciseness	Correlation, Classification
DialSummEval (Gao and Wan, 2022)	Model-generated summaries on dialogues, such as those in SAMSum (Gliwa et al., 2019), QMSum (Zhong et al., 2021), and MTSDialog (Ben Abacha et al., 2023a)	Coherence, Consistency, Fluency, Relevance	Correlation
DiaSumFact (Zhu et al., 2023)		Factual consistency	Classification
(Ben Abacha et al., 2023b)		Factual consistency	Correlation
Go Figure (Gabriel et al., 2021)	Model-generated and (transformed) reference summaries on news articles and dialogues	Faithfulness	Correlation
RoSE (Liu et al., 2023b)		Salience	Correlation
SummEdits (Laban et al., 2023)	Model-generated summaries, and LLM-edited reference summaries on diverse domains, such as news articles, scholarly articles, meeting transcripts, government reports, legal bills, etc.	Faithfulness	Classification
DiverSumm (Zhang et al., 2024a)		Faithfulness	Classification
(Ramprasad et al., 2024)		Faithfulness	Correlation

Table 1: A summary of recent benchmarks for meta-evaluating summarisation evaluation metrics.

Summary-level protocol calculates the correlation between X and Z on each summary first:

$$r_i, p = \mathbf{Corr}\left(\left[x_i^1, \cdot, x_i^J\right], \left[z_i^1, \cdot, z_i^J\right]\right), \quad (4)$$

and then apply an average operation to obtain the summary-level correlation coefficient:

$$r = \frac{1}{N}\sum_{i=1}^{N} r_i. \quad (5)$$

In addition to this common "correlation" perspective, recent studies focusing on evaluating the faithfulness of summaries also use classification or ranking protocols. That is, the generated summary (or more fine-grained elements, such as a sentence) is labelled by human annotators, for example as "faithful" or "unfaithful", and then automatic evaluation metrics are evaluated by whether they can predict accurately the label of a given summary (i.e., classification) or assigning a higher score to the faithful summary than the unfaithful summary (i.e., ranking).

We summarise recent benchmarks for meta-evaluating summarisation evaluation metrics in Table 1. A more detailed description of these benchmarks can be found in the Appendix C.

3 Discussion

We identify there are four critical decisions that must be made when we assess the automatic metric: (1) what source texts and summaries to use; (2) what quality dimensions to consider; (3) how to collect human judgements; and, (4) how to compare the automatic metric against human judgements. In this section, we discuss research trends and gaps around these four aspects.

3.1 Choosing Data to Annotate

Source texts From Table 1, we can see that most of the widely used meta-evaluation benchmarks use source texts from news summarisation datasets, followed by dialogue summarisation datasets. This is not ideal because, first, evaluation metrics tailored for evaluating news articles and summaries may not be portable to other domains due to the lack of respective resources. For example, QA-based evaluation metrics (Wang et al., 2020; Durmus et al., 2020) usually start from extracting named entities (e.g., person names) from source text and/or generated summary, around which questions are generated. However, entities of interest vary in different domains, and effective named entity recognition tools may not exist for specialised entity categories in niche domains, making these evaluation metrics hard to use. Secondly, the distribution of automatic evaluation scores usually differs across texts from various domains (Figure 1), and the generalisation ability of these evaluators, which are calibrated to the news domain, is underexplored (Laban et al., 2023). Finally, evaluation metrics usually show different correlation trends in different datasets, making their practical utility unclear. For example, Ramprasad et al. (2024) find that both QA-based and NLI-based evaluation metrics correlate well (Spearman's rank correlation coefficients of $0.45 \sim 0.59$) with human judgements on examples from news domain, but no correlation on biomedical domain (coefficients of $-0.03 \sim 0.11$).

Output summaries A common strategy for collecting summaries is assembling outputs from diverse summarisation systems, which are expected to cover different error types. For example, Clark et al. (2023) collect summaries from models of various sizes (e.g., 220M parameters of T5 (Raffel et al., 2020) and 540B parameters of PaLM (Aakanksha et al., 2024)) and employ both 1-shot in-context learning and fine-tuning approaches to generate the summaries. They also select both fully optimised and under-trained (i.e., trained for only 250 steps) checkpoints, ensuring differences in model quality.

Although these studies seek to diversify the summarisation systems, they often operate under a uniform summarisation formulation. In other words, the communicative goal and user preferences (e.g., the desired style and summary length) are disregarded when generating the summary. For example, Ramprasad et al. (2024) use the same prompt 'Article: [article]. Summarize the above article:' for generating summaries across domains. This simplified task formulation might be problematic when translating findings to build real-world summarisation applications. Summarisation involves compressing information in the source text by definition, and one key factor in this process is the compression ratio. Figure 2 shows that, under various constraints such as summary length (Koh et al., 2022), evaluation metrics may exhibit varying characteristics since generating shorter summaries (with a higher compression ratio) and evaluating these summaries is more challenging.

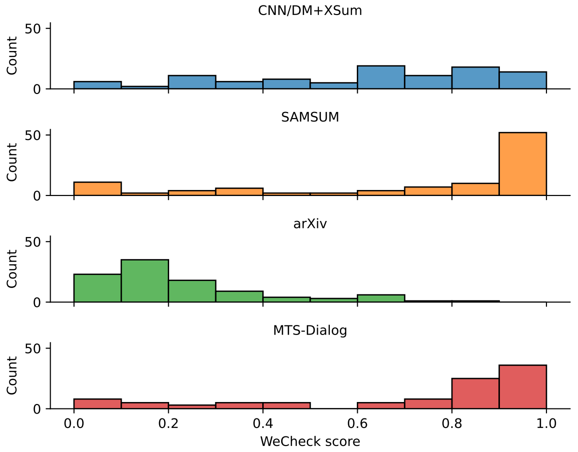

Figure 1: The distribution of consistency scores, measured using WeCheck (Wu et al., 2023), between source text and reference summary from different datasets. A score of 1 indicates a higher consistency level, while 0 indicates inconsistency. CNN/DM and XSUM datasets (Zhang et al., 2024b) include news articles, SAMSUM (Gliwa et al., 2019) messenger-like conversations, ARXIV (Cohan et al., 2018) scholarly articles, and MTSDIALOG (Ben Abacha et al., 2023a) from Doctor-Patient encounters.

Summary Because of the lack of meta-evaluation benchmarks covering various data distributions (i.e., source texts from different domains and output summaries from different systems under different task constraints), NLP practitioners may take the risk of overestimating the generalisation ability of automatic metrics (Chen et al., 2021). That is, practitioners may employ the top-performing evaluation metrics, e.g., for evaluating

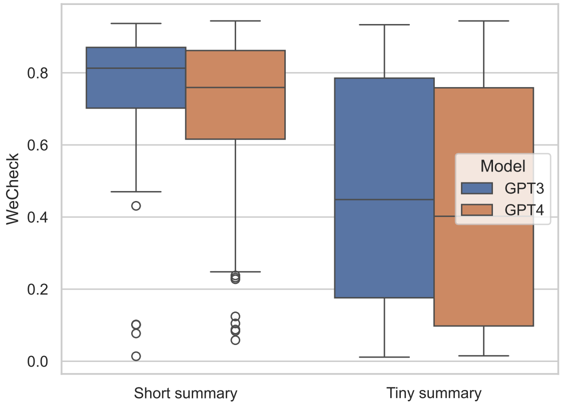

Figure 2: Evaluation results using WeCheck (Wu et al., 2023) on two tasks proposed in Multi-LexSum (Shen et al., 2022), where summaries are generated at different target levels of granularity: tiny (25 words, on average), and short (130 words). Prompts used to generate summaries can be found in Appendix Section A.

news summaries, and hope they work well for evaluating other types of summaries. To fill this gap, we call for building more diverse benchmarks that enable building more robust evaluation metrics and analysing the generalisation ability of existing evaluation metrics across different domains.

3.2 Defining Quality Dimensions

Mani (2001) divide the summarisation evaluation into two categories: intrinsic evaluation—testing the summarisation system as of itself—and extrinsic evaluation—testing based on how the generated summary affects the completion of some downstream tasks (e.g., efforts required to post-edit the generated summary to an acceptable, task-dependent state). We notice that most—if not all—recent benchmarks focus on quality dimension relating to the intrinsic evaluation but overlook the extrinsic evaluation.

From Table 1, we observe quality dimensions considered in recent benchmarks can be roughly grouped into two categories: (1) content quality, concerning to which extent the generated summary *accurately* reflects the most *important* information in the source text, and (2) language quality (e.g., coherency, fluency, comprehensibility) of the generated summary itself. We also notice that there is a shift in research focus towards content quality, especially the faithfulness of generated summaries.

Fonseca and Cohen (2024) argue that summarisation evaluation should consider the variability in communicative intentions. They choose three intentional aspects: conciseness (e.g., *Write a summary of the article above in 3 sentences*), narrative perspective (e.g., *Write in third person*), and keyword coverage (e.g., *Focus on the keywords: Thompson, sampling, sequential, variational*). They define *intention control metrics* to assess whether the generated summaries follow these intentions accurately. Zhang et al. (2024b) also point out that the summarisation evaluation should depend on the application scenarios and align with user values. They argue that, for example, the bullet point style summaries in CNN/DM (Nallapati et al., 2016) are rated by human annotators with low coherence scores; however, they may suffice for being displayed on news websites.

It is also worth noting that some quality dimensions are user-centric by nature, but most existing studies have overlooked the subjectivity of these dimensions. For example, when we define the readability of plain-language summaries of scientific articles (Goldsack et al., 2022), the end users' language and domain background should be taken into consideration. Another example is clinical conversation summarisation (Ben Abacha et al., 2023b); depending on whether the summary is provided for the clinicians or the patients, the same quality dimension (e.g, comprehensibility) should be defined differently.

Subtle differences behind the same term We observe a clear shift of research focus towards the content quality, especially faithfulness, of summarisation, mainly because recent LLM-based summarisation models have shown a remarkable capability to produce text of high language quality but still struggle with generating accurate content in a conditional-generation setting (Gao and Wan, 2022).

However, we also notice that different studies may investigate the same quality dimension following slightly different definitions, resulting in confusing conclusions. For example, Fabbri et al. (2020) define "consistency" as "whether the facts in the summary are consistent with the facts in the original article" but also "consider whether the summary does reproduce *all* facts accurately and does not makeup *unture* information". Honovich et al. (2022) define a text to be factually consistent to its grounding text (i.e., source text) "if all the factual information it conveys is consistent with the factual information conveyed by the grounding text" but "exclude personal and social statements". These subtle differences usually result in different judgements on the same summary due to "partial faithful" or "factual but not faithful" issues.

Summary We believe quality dimensions considered in the recent benchmarks are too narrow to reflect the various application scenarios where summarisation is used. Even worse, there is no census on the precise definition of these quality dimensions—different terms reflect the same underlying meaning and the same term refers to slightly different meanings—making the comparison against previous work difficult and unreliable.

3.3 Collecting Human Judgements

Who are expert annotators? Most previous studies, especially those that focus on news summarisation, refer to expert annotators as people who have experience in summarisation or NLP. Correspondingly, annotation guidelines are also heavily linguistic-oriented, for instance in their error categories and examples. For example, Pagnoni et al. (2021) collect human annotations based on a typology of factual errors, including "Relation Error", "Entity Error", "Circumstance Error", "Discourse Link Error", etc. Although this perspective can help developers understand the weaknesses of different summarisation models by examining the common errors these systems may generate, we argue that these errors may not always reflect real users' perspectives; instead, the real writers and readers of summaries should be more involved in the annotation process and the development of annotation guidelines.

Trade-off between annotation quality and cost Crowdsourcing is a common approach to reduce the time and cost associated with data annotation, though it often comes at the cost of sacrificing the reliability of the collected annotations. Most recent efforts that build meta-evaluation benchmarks rely on crowd annotators because crowd annotations can typically be collected quickly. In contrast, expert annotators may require significantly more time, even when fully dedicated to the annotation task. For example, Gao and Wan (2022) reported that they initially conducted the annotation via a crowd-sourcing platform and collected $7,000$ annotations from five different annotators in one day. In contrast, it took approximately 10 days to collect $4,200$ annotations from three student annotators.

Using LLMs as surrogate evaluators or combining LLM-as-evaluators with human evaluation to obtain an unbiased estimator with a lower cost than human evaluation alone is a promising but controversial research direction. Its effectiveness needs careful investigation as it depends not only on the correlation between the human and LLM-as-evaluator judgements but also on the choice of evaluation prompts (e.g., reference-free evaluation, pair-wise comparison, Likert survey) (Chaganty et al., 2018). Deutsch et al. (2022a) point out that when we use one generation model to evaluate another, they are biased against higher-quality outputs, including those written by humans. Liu et al. (2023a); Panickssery et al. (2024) also show that LLM-as-evaluators may have the problem of self-preference—they favour their own outputs or outputs from similar model families.

Ensuring annotation quality and detecting noisy annotations are then essential to building a reliable benchmark using crowd-sourcing or combining LLM-as-evaluators with human evaluation. However, we notice that only a limited number of quality control practices were commonly adopted in eliciting human annotations, such as filtering annotators based on their previous experience (Liu et al., 2023b), providing annotator training (Aharoni et al., 2023) and measuring inter-annotator agreement (Laban et al., 2023). Moreover, many studies overlook this issue and place blind trust in the collected data. For instance, Koto et al. (2022) found that only 7 out of 71 papers on summarisation human evaluation describe quality control mechanisms used.

Another overlooked practice is reporting failed attempts, which we believe can provide valuable insights to the following studies. For example, Gao and Wan (2022) hired 5 annotators using a crowd-sourcing platform to assess summaries generated from 14 different summarisation models on a Likert scale from 1 to 5. However, the model scores, which are calculated by averaging across 5 annotators on 100 summaries, are very close to each other (e.g., the averaged consistency score of the worst model is 3.206, whereas the best is 3.400), which they believe does not reflect reality.

The role of reference summary Intuitively, some quality dimensions can be assessed by reading the summary only. For example, Goldsack et al. (2022) instruct the annotators to rate *layness* (to what extent is the summary comprehensible to a non-expert) using a 1 to 5 Likert scale. However, these annotations usually suffer from inconsistency issues, as even the same annotator may make different assessments at different times.

Relative assessment, instead of direct assessment, is generally considered to improve agreement among annotators (Novikova et al., 2018). However, existing work uses reference summaries to aid human judgements mainly from a cost-saving consideration because annotators can rate the quality of a generated summary by comparing it against a short reference summary without reading a relatively long source text. Also, using a reference summary may reduce the annotation complexity for non-expert annotators. For example, Koto et al. (2022) argue that assessing *relevance*—the generated summary concisely captures all salient information—without a reference summary is difficult, as it requires annotators to implicitly construct their own summary of the source text.

However, the impact of reference summaries on human judgements and thus on meta-evaluation results is not well understood and examined. Regarding the same set of quality dimensions (*fluency*, *coherence*, *faithfulness* and *relevance*), Fabbri et al. (2020) provide annotators with summaries grouped in sets of 6 (i.e., 1 reference summary and 5 model-generated summaries), where the reference summary plays as an anchor between groups. But Zhuang et al. (2024) find that annotators tend to assign a lower score to the summary if it is shown along with a reference summary—even with a false reference summary. Automatic metric performance might also differ greatly depending on whether reference summaries are used during human annotations. For example, Liu et al. (2023b) find that reference-based metrics generally perform better when they are compared against human judgements collected using protocols with reference summary but can have negative correlations with those without using reference summary.

Human preferences vs quality judgement Instead of scoring summaries based on the description of quality dimensions, Goyal et al. (2022) adopt the approach of soliciting human preferences among summaries. However, this approach may be questionable when involving summaries generated using LLMs, which are usually pre-trained with human preference feedback. Liu et al. (2023b) point out that LLMs may have learned the prior preferences of human annotators but not necessarily captured the task-specific quality of summaries. Liu et al. designed two studies, asking human annotators: (a) to evaluate the summary without knowing the input text and (b) to evaluate if the summary covers the salient information of the input text. Results show that LLM-generated summaries received higher scores than human-written summaries under the first study, and the scores obtained from the first study are a good predictor of the results of the second study (Pearson's correlation of 0.926 between these two results). Zhang et al. (2024b); Shaib et al. (2024) also identify that annotators usually have their own consistent preference (e.g., based on summary length), when simply asked to rank the summaries.

Summary Given the costly nature of eliciting human judgements and the rapid pace of ongoing development in summarisation models, we believe there is an urgent need to standardise human evaluation practices. Developing a mechanism for producing reproducible human judgements over time and across different annotators (Khashabi et al., 2022) is paramount because it allows the collected resources to be reusable and easily extensible to new summarisation models. The resulting resources, which are more comprehensible, enable the development of effective and robust automatic metrics.

3.4 Comparing Automatic Metrics Against Human Judgements

Is a high correlation with human judgements enough to indicate the effectiveness of automatic metrics? A common way of reporting the effectiveness of automatic metrics is to tabulate the correlation between the results obtained using automatic metrics and human judgements, and metrics that achieve higher correlation are considered to be better (Fabbri et al., 2020; Ramprasad et al., 2024). However, Ernst et al. (2023) find that some evaluation metrics, although highly correlating with human judgements on a particular quality dimension, are, in fact, ineffective in measuring the considered dimension. For example, reference-based evaluation metrics correlate well with human judgements on *Fluency* and *Consistency* in the SUMMEVAL (Fabbri et al., 2020) benchmark but fail to detect even drastic summary corruptions, such as replacing all verbs with lemma form (resulting in ungrammatical summaries) and all person names with different names from the source text (resulting in unfaithful summaries).

One reason behind this phenomenon is that human judgements across quality dimensions may correlate with each other (Table 2 in Appendix B). Therefore, it is necessary to rule out the impact of confounding factors when comparing automatic metrics against human judgements for a particular quality dimension. Ernst et al. (2023) propose a bucketing-based approach where they divide all document-summary pairs into buckets where the human judgements of an anchor dimension have low variance; the correlations are calculated inside each bucket and then averaged with weights according to bucket size, resulting in more reliable meta-evaluation results for dimensions other than the anchor dimension.

Another reason is that most existing benchmarks include summaries generated from systems of varying quality. Therefore, high correlation is usually attributed to the capability of distinguishing between systems with large performance gaps. Deutsch et al. (2022b); Liu et al. (2023b); Shen et al. (2023) point out that discriminating between systems of similar quality is much more difficult than between systems of diverse quality, and a good metric should reliably indicate a difference in quality when a small difference in evaluation scores is observed. For example, Deutsch et al. found that the average improvement over baseline models reported in recent papers on the CNN/DM (Nallapati et al., 2016) dataset was ROUGE-1 score of 0.5. However, the correlation of ROUGE-1 to human judgements is near 0 when ranking systems whose evaluation scores are so close. On the other hand, a large gap (e.g., 5-10) of ROUGE scores does correctly rank system pairs, enabling ROUGE to achieve moderately strong correlations on standard benchmarks.

Statistical Power concerns the chance a *significant* difference (e.g., evaluation metrics score differently in a meta-evaluation benchmark) will be observed, given there is a real difference (i.e., genuinely different evaluation metrics) (Card et al., 2020). Deutsch et al. (2021b) find that high uncertainty (large confidence intervals) exists when evaluating automatic metrics using existing benchmarks. This is also observed in human evaluation of similar-performing systems (Liu et al., 2023b). Although increasing the dataset's sample size (but requiring a significant human effort) can effectively raise statistical power (Shaib et al., 2024), other cheap alternatives are needed. For example, Deutsch et al. (2022b) propose to calculate automatic scores on a much larger set instead of only the subset of summaries judged by humans.

Summary We argue that assessing the effectiveness of automatic metrics can be conducted in multiple stages, each requiring different levels of human annotation effort. First, evaluation metrics should be tested on their effectiveness in detecting significant errors, e.g., corruptions in human-written summaries (Gabriel et al., 2021; Chen et al., 2021; Ernst et al., 2023). Secondly, they can be meta-evaluated against existing human judgements on summaries from systems of varying quality. Thirdly, human judgements should be constantly gathered on summaries generated using state-of-the-art systems, presumably of closing quality (Peyrard, 2019) and automatic metrics should be tested on discriminating these systems. Finally, metrics should be tested against reproducing human preferences between pairs of summaries (i.e., summary-level effectiveness) and the capability of identifying more fine-grained problems (Chen et al., 2021).

4 Related Work

Similar to summarisation, other natural language generation tasks, such as machine translation (MT), dialogue, and data-to-text generation, also have a long history of employing automatic evaluation metrics, such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), to assess the quality of machine-generated texts. Assessing the effectiveness and reliability of these automatic metrics is also an active research area, and regular shared tasks (e.g., WMT Metrics Shared Task[1]) are organised to encourage researchers to explore the strengths and weaknesses of automatic metrics. Unfortunately, similar efforts to meta-evaluate summarisation evaluation metrics were unsustained, partially due to the complexity of the summarisation task itself. As observed by Graham (2015), although there are obvious parallels between summarisation and machine translation (MT), methodologies applied to meta-evaluate MT metrics have not been well explored in summarisation.

With the advancement of large-scale generative models, evaluating text generated by LLMs and meta-evaluate corresponding evaluation metrics have also attracted significant interests (Gehrmann

[1] https://www2.statmt.org/wmt24/metrics-task.html

et al., 2021; Zhao et al., 2024; Li et al., 2023; Pal et al., 2023; Mishra et al., 2024). These studies usually concern similar quality dimensions as those in the summarisation field, and have a similar desire to find cost-effective ways to collect human judgements.

5 Conclusions and Recommendations

In this position paper, we critically examine the practices of meta-evaluating summarisation evaluation metrics in the literature. We identify several avenues in the field that can be further improved, regarding: *choosing data to annotate*, *defining quality dimensions*, *collecting human judgements*, and *comparing automatic metrics against human judgements*.

For practitioners aiming to assess the effectiveness of automatic metrics for their particular use case, we suggest starting by considering the role a summarisation system plays in the real-world workflow. This includes identifying the readers of the generated summaries, understanding what information they seek, and what decisions they might make after reading the summary. Once we have a clear picture of this, we can create document-summary pairs that meet requirements and focus on the quality dimensions that end-users value most. Human judgements can be collected from real end-users regarding both their perceived quality of the summary and the effect of these summaries on the actual downstream tasks they perform. Finally, automatic evaluations can be assessed depending on the purposes of the evaluation, such as determining which summarisation system is better (system-level correlation is informative), choosing the best summary from multiple candidates (summary-level correlation/ranking), and detecting problematic summaries (binary classification).

For researchers who aim to develop meta-evaluation resources and novel evaluation metrics, we believe it is time to build more diverse benchmarks using data sampled from different domains and considering various summarisation constraints. That is to say, the generality of evaluation metrics should be tested to mitigate the risk of overestimating the effectiveness of automatic metrics across domains and applications (Gabriel et al., 2021). Secondly, we believe there is an urgent need to standardise human evaluation practices to ensure reproducible human judgements over time and, more importantly, to make the collected resources extensible to new summarisation models. We recommend some best (basic) practices: (1) being aware of previous work and reusing previous resources (taxonomy, guideline, interface, etc.) whenever possible (Tang et al., 2023); (2) adopting quality controls, such as training annotators to make sure they understand the annotation task, filtering out unqualified annotators and their annotations, etc (Koto et al., 2022); (3) documenting the creation process (e.g., preprocessing, annotating) and recommended uses (Gebru et al., 2021). Finally, we argue that claims on the usefulness of evaluation metrics should be made based on comprehensive and reliable assessment under various usage scenarios, such as detecting summaries with significant errors, distinguishing summary systems of closing quality, or identifying more fine-grained issues in the generated summary.

Limitations

The main limitation of this study is that we did not contribute any new resources or procedures for meta-evaluation. This study states our opinions based on our educated review of the current literature.

References

Chowdhery Aakanksha, Narang Sharan, Devlin Jacob, Bosma Maarten, Mishra Gaurav, Roberts Adam, Barham Paul, Chung Hyung Won, Sutton Charles, Gehrmann Sebastian, Schuh Parker, Shi Kensen, Tsvyashchenko Sashank, Maynez Joshua, Rao Abhishek, Barnes Parker, Tay Yi, Shazeer Noam, Prabhakaran Vinodkumar, Reif Emily, Du Nan, Hutchinson Ben, Pope Reiner, Bradbury James, Austin Jacob, Isard Michael, Gur-Ari Guy, Yin Pengcheng, Duke Toju, Levskaya Anselm, Ghemawat Sanjay, Dev Sunipa, Michalewski Henryk, Garcia Xavier, Misra Vedant, Robinson Kevin, Fedus Liam, Zhou Denny, Ippolito Daphne, Luan David, Lim Hyeontaek, Zoph Barret, Spiridonov Alexander, Sepassi Ryan, Dohan David, Agrawal Shivani, Omernick Mark, M. Dai Andrew, Pillai Thanumalayan Sankaranarayana, Pellat Marie, Lewkowycz Aitor, Moreira Erica, Child Rewon, Polozov Oleksandr, Lee Katherine, Zhou Zongwei, Wang Xuezhi, Saeta Brennan, Diaz Mark, Firat Orhan, Catasta Michele, Wei Jason, Meier-Hellstern Kathy, Eck Douglas, Dean Jeff, Petrov Slav, and Fiedel Noah. 2024. PaLM: scaling language modeling with pathways. *JMLR*, 24.

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual Summarization with Factual Consistency Evaluation. In *Findings of ACL*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023a. An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. In *EACL*.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023b. An Investigation of Evaluation Methods in Automatic Medical Note Generation. In *Findings of ACL*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating Evaluation in Text Summarization. In *EMNLP*.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With Little Power Comes Great Responsibility. In *EMNLP*.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *ACL*.

Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara, Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan, Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi, and Alexander Gray. 2022. X-FACTOR: A Cross-metric Evaluation of Factual Correctness in Abstractive Summarization. In *EMNLP*.

Yiran Chen, Pengfei Liu, and Xipeng Qiu. 2021. Are Factuality Checkers Reliable? Adversarial Meta-evaluation of Factuality in Summarization. In *Findings of EMNLP*.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A Multilingual, Multifaceted Dataset for Summarization Evaluation. In *EMNLP*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *NAACL*.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021a. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. *TACL*, 9.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021b. A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods. *TACL*, 9.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022a. On the Limitations of Reference-Free Evaluations of Generated Text. In *EMNLP*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022b. Re-Examining System-Level Correlations of Automatic Summarization Evaluation Metrics. In *NAACL*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In *ACL*.

Ori Ernst, Ori Shapira, Ido Dagan, and Ran Levy. 2023. Re-Examining Summarization Evaluation across Multiple Quality Criteria. In *Findings of EMNLP*.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Richard Socher, and Dragomir R. Radev. 2020. SummEval: Re-evaluating Summarization Evaluation. *TACL*, 9.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In *ACL*.

Marcio Fonseca and Shay B Cohen. 2024. Can Large Language Model Summarizers Adapt to Diverse Scientific Communication Goals? In *Findings of ACL*.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A Meta Evaluation of Factuality in Summarization. In *Findings of ACL*.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting Summarization Evaluation for Dialogues. In *NAACL*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64.

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 *Workshop on Natural Language Generation, Evaluation, and Metrics*.

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

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In *Workshop on New Frontiers in Summarization*.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the BioLaySumm 2023 Shared Task on Lay Summarization of Biomedical Research Articles. In *Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature. In *EMNLP*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News Summarization and Evaluation in the Era of GPT-3. *arXiv*, 2209.12356.

Yvette Graham. 2015. Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. In *EMNLP*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In *Findings of ACL-IJCNLP*.

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 *NAACL*.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What Have We Achieved on Text Summarization? In *EMNLP*.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead. In *Workshop on Human Evaluation of NLP Systems*.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel Weld. 2022. GENIE: Toward Reproducible and Standardized Human Evaluation for Text Generation. In *EMNLP*.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How Far are We from Robust Long Abstractive Summarization? In *EMNLP*.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. FFCI: A Framework for Interpretable Automatic Evaluation of Summarization. *JAIR*, 73.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *EMNLP*.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM Ability at Factual Reasoning Through The Lens of Summarization. In *EMNLP*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *TACL*, 10.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *EMNLP*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2023a. LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores. In *Findings of ACL*.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq R. Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir R. Radev. 2023b. Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation. In *ACL*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing Order to Abstractive Summarization. In *ACL*.

Liang Ma, Shuyang Cao, Robert L. Logan IV, Di Lu, Shihao Ran, Ke Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. BUMP: A Benchmark of Unfaithful Minimal Pairs for Meta-Evaluation of Faithfulness Metrics. In *ACL*.

Inderjeet Mani. 2001. Summarization Evaluation: An Overview. In *Workshop on Research in Chinese and Japanese Text Retrieval and Text Summarization*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *COLM*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *CoNLL*.

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 *EMNLP*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable Human Ratings for Natural Language Generation. In *NAACL*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In *NAACL*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical Domain Hallucination Test for Large Language Models. In *CoNLL*.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In *NeurIPS*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

Maxime Peyrard. 2019. Studying Summarization Evaluation Metrics in the Appropriate Scoring Range. In *ACL*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21.

Sanjana Ramprasad, Kundan Krishna, Zachary C Lipton, and Byron C Wallace. 2024. Evaluating the Factuality of Zero-shot Summarizers Across Varied Domains. In *EACL*.

Peter Rankel, John Conroy, Eric Slud, and Dianne O'Leary. 2011. Ranking Human and Machine Summarization Systems. In *EMNLP*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *ACL*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The Multilingual Summarization Corpus. In *EMNLP*.

Chantal Shaib, Joe Barrow, Alexa F Siu, Byron C Wallace, and Ani Nenkova. 2024. How Much Annotation is Needed to Compare Summarization Models? In *Workshop on Bridging Human–Computer Interaction and Natural Language Processing*.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. In *Findings of EMNLP*.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities. In *NeurIPS*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *NeurIPS*.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the Factual Consistency of Large Language Models Through News Summarization. In *Findings of ACL*.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors. In *ACL*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *ACL*.

Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Sujian Li, and Yajuan Lyu. 2023. WeCheck: Strong Factual Consistency Checker via Weakly Supervised Learning. In *ACL*.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024a. Fine-Grained Natural Language Inference Based Faithfulness Evaluation for Diverse Summarisation Tasks. In *EACL*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024b. Benchmarking Large Language Models for News Summarization. *TACL*, 12.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020b. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *ACL*.

Yiran Zhao, Jinghan Zhang, I. Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2024. Felm: Benchmarking factuality evaluation of large language models. In *NeurIPS*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *NAACL*.

Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and Detecting Fine-grained Factual Errors for Dialogue Summarization. In *ACL*.

Haojie Zhuang, Wei Emma Zhang, Leon Xie, Weitong Chen, Jian Yang, and Quan Z Sheng. 2024. Automatic, Meta and Human Evaluation for Multimodal Summarization with Multimodal Output. In *NAACL*.

A Implementation Details

A.1 Prompts used to generate summaries on tasks proposed in MULTI-LEXSUM

- Short summary (GPT-3.5)

  System: As a junior legal intern, please craft a summary (approximately 130 words) for the given legal case.

  User: [article]

- Tiny summary (GPT-3.5)

  System: As a junior legal intern, please craft a summary (approximately 25 words) for the given legal case.

  User: [article]

- Short summary (GPT-4)

  System: As a senior legal professional, please craft a summary (approximately 130 words) for the given legal case.

  User: [article]

- Tiny summary (GPT-4)

  System: As a senior legal professional, please craft a summary (approximately 25 words) for the given legal case.

  User: [article]

B Additional Results

Table 2 show the correlation between different quality dimensions within the same annotator group and across different groups for the same dimension.

	Coher.	Faith.	Fluen.	Rele.	Expert
Expert annotators					
Coherence	-	0.300	0.544	0.700	**0.877**
Faithfulness	0.300	-	0.594	0.500	**0.900**
Fluency	0.544	0.594	-	0.745	**0.810**
Relevance	0.700	0.500	0.745	-	**0.857**
Crowd annotators					
Coherence	-	0.310	**0.500**	0.393	-0.083
Faithfulness	0.310	-	**0.343**	0.168	0.059
Fluency	**0.500**	0.343	-	0.326	0.142
Relevance	**0.393**	0.168	0.326	-	-0.159

Table 2: System-level Kendall's τ correlation coefficients between different quality dimensions within the same annotator group, and correlation coefficients between different annotator groups for the same quality dimension. underline: the correlation coefficient is significant ($p \leq 0.05$). The human judgements are from SUMMEVAL (Fabbri et al., 2020).

C Meta-evaluation Benchmarks

SUMMEVAL Fabbri et al. (2020) assembled a collection of summaries generated by 16 models trained on the CNN/DM (Nallapati et al., 2016) dataset and collect human judgements from 3 expert judges and 5 crowd-source workers. Judges were asked to evaluate the summaries along four dimensions: relevance (concerning the selection of important content), consistency (concerning factual alignment between the summary and the source), fluency (concerning the quality of individual sentences), and coherence (concerning the collective quality of all sentences).

REALSUMM Bhandari et al. (2020) released a dataset of human judgements on the relevance of summaries collected from 25 neural summarization systems. Bhandari et al. create Semantic Content Units (SCUs) for each reference summary and then hire crowd workers to annotate each generated summary, determining whether each SCU can be inferred from the generated summary.

FRANK Pagnoni et al. (2021) devise a typology of factual errors (e.g., predicate errors, entity errors, circumstance errors, etc.) and use it to collect human annotations of generated summaries for the CNN/DM (Nallapati et al., 2016) and XSUM (Narayan et al., 2018) datasets. They conduct the annotation task on the Amazon Mechanical Turk platform and found a nearly perfect agreement (a Cohen Kappa of 0.86) between the majority class of the three crowd annotators and one expert annotator on 20 summaries.

GO FIGURE Gabriel et al. (2021) introduce a meta-evaluation framework for evaluating factuality evaluation metrics. Gabriel et al. build one diagnostic dataset that consists of transformed reference summaries with simulated factuality errors (i.e., pronoun entity errors, verb tense or negation errors, intrinsic entity errors, extrinsic entity errors, sentiment errors, false quotes). They also use fine-tuned T5 summarization models to generate summaries and annotate them for fine-grained factual errors based on the above-mentioned error types.

DIALSUMMEVAL Gao and Wan (2022) sample 100 dialogues from the SAMSUM (Gliwa et al., 2019) test set and evaluate the summaries generated by 14 summarization models. Three college students fluent in English were recruited to assess the relevance, consistency, fluency and coherence quality of generated summaries.

BUMP Ma et al. (2023) introduce a dataset of 889 summary pairs, where a single error is introduced to a reference summary from the CNN/DM (Nallapati et al., 2016) dataset to produce an unfaithful summary. Ma et al. define a taxonomy of seven unfaithful error types (i.e., intrinsic/extrinsic predicate error, intrinsic/extrinsic entity error, intrinsic/extrinsic circumstance error, and coreference error) and instruct annotators to introduce errors of a specific type.

ROSE Liu et al. (2023b) propose a 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. Using the protocol, Liu et al. curate a large human evaluation dataset consisting of 22,000 summary-level annotations over 28 systems on samples from CNN/DM (Nallapati et al., 2016), XSUM (Narayan et al., 2018), and SAMSUM (Gliwa et al., 2019).

SEAHORSE Clark et al. (2023) collect annotations along 6 dimensions: comprehensible (read and understood by the rater), repetition (free of unnecessarily repeated information), grammar (grammatically correct), attribution (fully attributable to the source article), main ideas (captures the main ideas of the source article), and, conciseness (concisely represents the information in the source article). Annotators can answer 'Yes,' 'No,' or 'Unsure' to the first three questions given only the summary, and the last three questions given both the article and the summary. Their annotations provide both a benchmark for meta-evaluation but also a resource for training learning-based evaluation metrics.

SUMMEDITS Laban et al. (2023) propose a new protocol for creating inconsistency detection benchmarks. First, they manually verify the factual consistency of a small set of seed summaries. Then, they use LLMs to generate numerous edited versions (e.g., via entity modification, antonym swap, hallucinated fact insertion, and negation insertion) of these consistent seed summaries. Finally, human annotators determine whether each edit introduces a factual inconsistency. Laban et al. implement the protocol on ten diverse textual domains, including the legal, dialogue, academic, financial, and sales domains.

FIB Tam et al. (2023) propose a factual inconsistency benchmark, where each example consists of a document and two summaries (one factually consistent summary and one factually inconsistent summary). For factually consistent summaries, they consider reference summaries from CNN/DM (Nallapati et al., 2016) and XSUM (Narayan et al., 2018) and manually fix these factually inconsistent reference summaries using minimal edits. They also manually choose factually inconsistent summaries from model-generate summaries.