Automated Metrics for Medical Multi-Document Summarization Disagree with Human Evaluations

Lucy Lu Wang[1,2] **Yulia Otmakhova**[3] **Jay DeYoung**[4] **Thinh Hung Truong**[3]
Bailey E. Kuehl[2] **Erin Bransom**[2] **Byron C. Wallace**[4]

[1]University of Washington [2]Allen Institute for AI [3]University of Melbourne
[4]Northeastern University

lucylw@uw.edu, {yotmakhova, hungthinht}@student.unimelb.edu.au
{deyoung.j, b.wallace}@northeastern.edu

Abstract

Evaluating multi-document summarization (MDS) quality is difficult. This is especially true in the case of MDS for biomedical literature reviews, where models must synthesize contradicting evidence reported across different documents. Prior work has shown that rather than performing the task, models may exploit shortcuts that are difficult to detect using standard n-gram similarity metrics such as ROUGE. Better automated evaluation metrics are needed, but few resources exist to assess metrics when they are proposed. Therefore, we introduce a dataset of human-assessed summary quality facets and pairwise preferences to encourage and support the development of better automated evaluation methods for literature review MDS. We take advantage of community submissions to the Multi-document Summarization for Literature Review (MSLR) shared task to compile a diverse and representative sample of generated summaries. We analyze how automated summarization evaluation metrics correlate with lexical features of generated summaries, to other automated metrics including several we propose in this work, and to aspects of human-assessed summary quality. We find that not only do automated metrics fail to capture aspects of quality as assessed by humans, in many cases the system rankings produced by these metrics are anti-correlated with rankings according to human annotators.[1]

1 Introduction

Multi-document summarization (MDS) requires models to summarize key points across a set of related documents. Variants of this task have drawn significant attention in recent years, with the introduction of datasets in domains like newswire (Fabbri et al., 2019), Wikipedia (Gholipour Ghalandari et al., 2020), science (Lu et al., 2020), medical literature reviews (DeYoung et al., 2021; Wallace et al., 2020), and law (Shen et al., 2022); and substantial methodological work to design model architectures tailored to this task (Xiao et al., 2022; Pasunuru et al., 2021; Liu and Lapata, 2019).

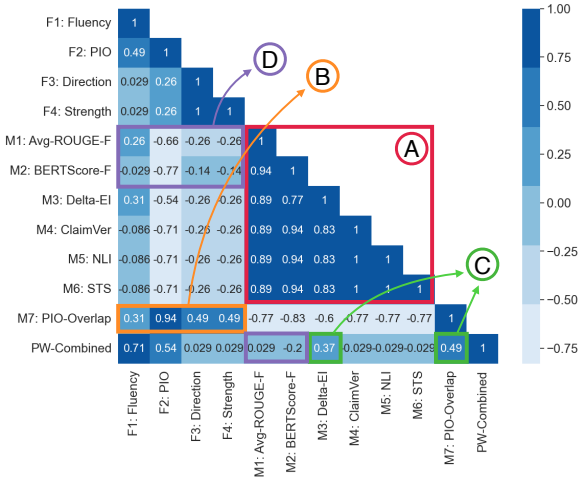

Figure 1: Spearman correlations between rankings produced by human-assessed quality facets (F1-F4), automated metrics (M1-M7), and combined pairwise system rankings (PW-combined) on the Cochrane MSLR dataset. Rankings from automated metrics are highly correlated as a group except for PIO-Overlap (A). PIO-Overlap rankings are strongly correlated with rankings from human-assessed facets, especially PIO agreement (B). Metrics most strongly associated with PW-Combined rankings are Delta-EI and PIO-Overlap (C). Rankings from commonly reported automated metrics like ROUGE and BERTScore are not correlated or *anti*-correlated with human-assessed system rankings (D).

In this work, we focus on MDS for literature reviews (MSLR), a challenging variant of the task in which one attempts to synthesize all evidence on a given topic. When manually performed, such reviews usually take teams of experts many months to complete. Good review summaries aggregate the results of different studies into a coherent passage, while the evidence presented in the input studies will often be in conflict (Wallace et al., 2020; DeYoung et al., 2021; Wadden et al., 2022), complicat-

[1]Dataset and analysis are available at https://github.com/allenai/mslr-annotated-dataset.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9871–9889
July 9-14, 2023

ing the synthesis task.[2]

Evaluating conditional text generation models is notoriously difficult, impeding progress in the field. Prior work on summarization evaluation has proposed various lexical and modeling-based approaches to assess generation quality, but these metrics predominately use correlation with human-assessed quality facets over relatively small numbers of examples to demonstrate utility (Fabbri et al., 2021; Wang et al., 2020; Deutsch and Roth, 2020; Yuan et al., 2021). This limitation of current metric evaluation implies that existing automated measures may not generalize well. Further, evaluation in the multi-document setting adds additional complexity, e.g., prior work has shown that MDS models may sometimes exploit shortcuts that do not reflect as detectable changes in automated metrics (Wolhandler et al., 2022; Giorgi et al., 2022a).

To address these challenges, we collect human annotations to evaluate current models and to support automated metrics development for the medical MDS task. We construct a dataset of such evaluations using public submissions from the 2022 MSLR shared task on literature review MDS.[3] Selecting top-performing models, we label the summary quality of a sample of these models' outputs on the Cochrane subtask (Wallace et al., 2020). As part of our analysis, we compare system rankings produced by automated metrics and human evaluations. Strikingly, our results highlight consistent and significant disagreements between automated metrics and humans, motivating the need for better automated evaluation metrics in this domain.

We contribute the following:

- A dataset of summaries and quality annotations on participant submissions to the MSLR shared task. We include human annotations for 6 models on 8 individual quality facets (§3.2) and pairwise preferences provided by five raters (§3.3).
- An analysis of lexical features among inputs, generated, and target summaries (§4), showing a large amount of undesirable copying behavior.
- An analysis of correlations between automated evaluation metrics and human-assessed quality (§5), and the differences in system rankings produced by automated metrics versus human evaluation (§6). We propose several novel evaluation metrics based on desired features of MSLR summaries (§5). We find that system rankings derived from commonly reported automated metrics are *not* correlated or even *anti*-correlated with rankings produced by human assessments of quality, though some of the metrics we propose demonstrate promise in capturing certain quality facets.

2 Background

The MSLR shared task was introduced to bring attention to the challenging task of MDS for literature reviews. The shared task comprised two subtasks, based on the Cochrane (Wallace et al., 2020) and MSˆ2 (DeYoung et al., 2021) datasets. The Cochrane dataset consists of 4.6K reviews from the Cochrane database of systematic reviews. Inputs are abstracts of papers cited by the review and target summaries are the *Authors' Conclusions* subsections of review abstracts. The MSˆ2 dataset includes 20K reviews and is semi-automatically constructed from biomedical literature reviews indexed by PubMed. We refer the reader to the original publications for details concerning dataset construction (Wallace et al., 2020; DeYoung et al., 2021).

Shared task organizers provided training and validation splits for both datasets, and solicited model submissions to two public leaderboards, where models were evaluated on a hidden test split. Models were ranked on the leaderboard using ROUGE (-1, -2, -L; Lin 2004), BERTScore (Zhang et al., 2020a), and Delta-EI (DeYoung et al., 2021; Wallace et al., 2020), a metric based on evidence inference (Lehman et al., 2019) classifications.

3 Dataset

We construct our dataset from system submissions to the Cochrane subtask leaderboard for the 2022 MSLR shared task (provided to us by task organizers). We only sample from the Cochrane subtask due to the greater number and variety of successful submissions. We include all summaries from the leaderboard, though we only perform human evaluation on summaries generated by 6 models (discussion in §3.1). We define and apply two human evaluation protocols to a sample of summaries from these 6 systems. The first (§3.2) is a facet-based evaluation derived from the analysis conducted in Otmakhova et al. (2022b) and the second (§3.3) is a pairwise preference assessment.

[2]Indeed, reviews conducted by different teams may themselves conflict (Ioannidis, 2016), reflecting the inherent difficulty of the task; however this may owe to differing methods of selecting input studies, a complication we ignore here, though which has been explored in recent work (Giorgi et al., 2022a).

[3]https://github.com/allenai/mslr-shared-task

3.1 MDS systems

We perform human evaluation on the outputs of 6 MDS systems. Five of these are community submissions to the MSLR-Cochrane leaderboard,[4] while a sixth is a baseline system (BART-Cochrane) included for reference. These systems represent different Transformer model architectures (BART, BART-large, Longformer, BigBird), input selection strategies (Shinde et al., 2022), and differential representation/attention on input tokens (Otmakhova et al., 2022a; DeYoung et al., 2021). We exclude some systems from human evaluation due to poor summary quality (disfluent) or being baselines. We briefly describe our 6 systems below.

ITTC-1 / ITTC-2 Otmakhova et al. (2022a) fine-tuned PRIMERA (Xiao et al., 2022) for the Cochrane subtask and exploited the use of global attention to highlight special entities and aggregate them across documents. We include two settings from the leaderboard, one that adds global attention to special entity marker tokens (ITTC-1) and one that adds global attention to entity spans (ITTC-2).

BART-large Tangsali et al. (2022) fine-tuned BART-large (Lewis et al., 2020) for the subtask.

SciSpace Shinde et al. (2022) defined an *extract-then-summarize* approach, combining BERT-based extraction of salient sentences from input documents with a BigBird PEGASUS-based summarization model (Zaheer et al., 2020).

LED-base-16k Giorgi et al. (2022b) fine-tuned Longformer Encoder-Decoder (Beltagy et al., 2020) for the Cochrane subtask following a similar protocol described in Xiao et al. (2022).

BART (baseline) The baseline follows the protocol in DeYoung et al. (2021) to fine-tune BART (Lewis et al., 2020) for the Cochrane subtask.
Model rankings originally reported on the MSLR-Cochrane leaderboard are provided in Table 1.

3.2 Facet-based Human Evaluation

We adapt a facet-based human evaluation procedure from the analysis in Otmakhova et al. (2022b). In their work, the authors analyzed baseline model outputs from MSˆ2 (DeYoung et al., 2021) with respect to fluency, PIO alignment, evidence direction, and modality (or strength of claim). PIO stands for Population (who was studied? e.g. women with gestational diabetes), Intervention (what was studied? e.g. metformin), and Outcome (what was measured? e.g. blood pressure), and is a standard framework for structuring clinical research questions (Huang et al., 2006). These are important elements that *must* align between generated and target summaries for the former to be considered accurate. Evidence direction describes the effect (or lack thereof) that is supported by evidence (e.g., the treatment shows a positive effect, no effect, or a negative effect, comparatively). The strength of the claim indicates how much evidence or how strong the evidence associated with the effect might be.

We derive 8 questions based on this analysis:

1. *Fluency*: if the generated summary is fluent
2. *Population*: whether the population in the generated and target summaries agree
3. *Intervention*: as above for intervention
4. *Outcome*: as above for outcome
5. *Effect-target*: effect direction in the target
6. *Effect-generated*: effect direction in the generated summary
7. *Strength-target*: strength of claim in the target
8. *Strength-generated*: strength of claim in the generated summary

Of the 470 reviews in the Cochrane test set, we sample 100 reviews per system for facet annotations (600 summaries in total). For 50 reviews, we fully annotate all summaries from the 6 systems (the overlapping set); for the other 50 reviews per system, we sample randomly from among the remaining reviews for each system (the random set). All together, at least one system's outputs are annotated for 274 reviews in the test set. We elect for this sampling strategy to balance thoroughness (having sufficient data points to make direct comparisons between systems) and coverage (having annotations across more review topics).

For each sampled instance, we show annotators a pair of (target, generated) summaries from a review and ask them to answer 8 questions regarding these (details in App. A). A sample of 10 reviews from the overlapping set (60 summary pairs) and 10 from the random set (10 summary pairs) are annotated by two annotators. We compute inter-annotator agreement from these and report Cohen's Kappa and agreement proportions for all eight facets in Table 2. Several facets have lower agreement (Population, Outcome, and Strength-target), though most disagreements are between similar classes (e.g. par-

[4] https://leaderboard.allenai.org/mslr-cochrane/

System	ROUGE*	BERTS.	ΔEI	ClaimV.	NLI	STS	PIO-Over.	Flu.	PIO	Dir.	Str.	PW-Comb.
ITTC-1	5 (4)	5 (2)	4 (6)	4	4	4	1	3	1	3	3	1
ITTC-2	1 (2)	2 (1)	1 (2)	2	2	2	5	1	4	6	6	2
BART-large	3 (6)	3 (5)	2 (4)	3	3	3	4	4	5	2	2	3
LED-base-16k	4 (3)	4 (3)	5 (5)	5	5	5	2	2	2	1	1	4
SciSpace	2 (1)	1 (6)	3 (3)	1	1	1	6	6	6	4	4	6
BART (baseline)	6 (5)	6 (4)	6 (1)	6	6	6	3	5	3	5	5	5

Table 1: System rankings based on automated metrics and human evaluation (best in green). Original system ranks from the MSLR leaderboard as assessed based on ROUGE-L, BERTScore, and Delta-EI are provided in parentheses. The ranks in this table are produced over subsamples of reviews from the Cochrane test split (and macro-averaged for ROUGE and BERTScore), causing ranks to differ from leaderboard rankings.
*Ranking for ROUGE is based on Avg-ROUGE-F, while leaderboard rank is based on ROUGE-L.

Question	Classes	κ	Agreement
Fluency	3	0.52	0.87
Population	4	0.33	0.56
Intervention	4	0.60	0.77
Outcome	4	0.24	0.36
Effect-target	4	0.85	0.90
Effect-generated	4	0.78	0.90
Strength-target	4	0.30	0.54
Strength-generated	4	0.77	0.90

Table 2: Inter-annotator agreement between experts on facets (Cohen's κ and proportion of agreement).

tial agree vs. agree); more on this in App. A.

Two annotators with undergraduate biomedical training annotated these samples. We arrived at the final annotation protocol following two rounds of pilot annotations on samples from the MSˆ2 dataset and discussing among authors to resolve disagreements and achieve consensus.

3.3 Pairwise Human Evaluation

We perform pairwise comparisons to elicit human preferences between system-generated summaries and to study how facet-based quality maps to holistic summary quality.

We sample pairs of system generations from our dataset, half from the overlapping set of reviews annotated for facet evaluations, and half from other reviews. A different subsample of these pairwise comparisons is provided to each of 5 raters, who are asked to complete up to 100 judgments each. For each comparison, the annotator is given the target summary, the system A summary, the system B summary, and asked "Which of A or B more accurately reflects the content of the target summary?" where the options are A, B, or Neither. All annotators are knowledgable in BioNLP and one annotator has biomedical training. Four annotators completed 100 pairwise comparisons; a fifth completed 50 comparisons.

We first determine system rankings per individual annotator. To tally annotations: if A is preferred over B, system A gets 1 point; if B over A, system B gets 1 point; if Neither is preferred, neither system gets a point. Systems are ranked by total points; tied systems receive the same ranking. To determine a combined ranking based on the preferences of all 5 annotators, we adopt the Borda count (Emerson, 2013), a ranked choice vote counting method that maximizes the probability of selecting the Condorcet winner.[5] In this method, for each annotator (voter), we award each system the number of points corresponding to the number of systems ranked below it, e.g., for a set of systems ranked 1-6, the rank 1 system receives 5 points, the rank 2 system 4 points, and so on. System rankings resulting from the Borda count are shown in Table 1 under Pairwise-Combined.

We perform bootstrapping over each annotator's pairwise annotations to estimate the error of the overall system rankings. We resample each individual's pairwise preferences with replacement and compute a new combined ranking. Over 10000 bootstrap samples, the average Spearman ρ of the resampled rankings against the initial rankings is 0.716 (s/d = 0.197).

3.4 Dataset Statistics

Our final dataset consists of 4658 summaries generated by 10 systems over 470 review instances from MSLR-Cochrane. Of these summaries, 597 from 6 systems are annotated on 8 quality facets. We also include 452 pairwise comparisons from five annotators. In addition to annotations, we compute and include automated metrics for each generated summary to facilitate analysis (more in §5).

[5]The Condorcet winner is the candidate that would win a head-to-head election against each of the other candidates assuming a plurality vote.

System	Synthesis	Input Match
Targets	0.48	-
ITTC1	0.46	0.26
ITTC2	0.45	0.15
BART-Large	0.41	0.31
LED-Base-16K	0.45	0.36
SciSpace	0.44	0.48
BART (baseline)	0.44	0.38

Table 3: Results of summary vs. input lexical analysis.

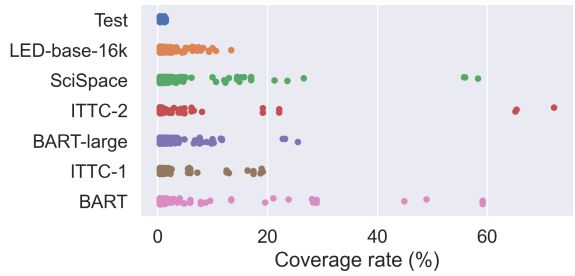

Figure 2: Percent of summaries in which each self-repeating 7-gram appears (Test = target summaries).

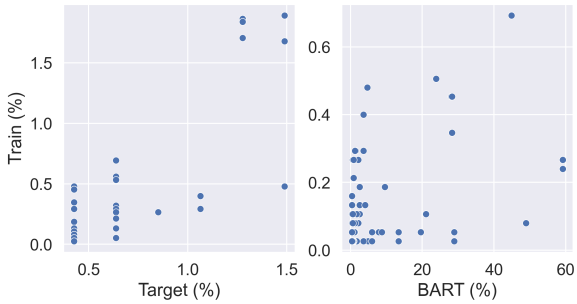

Figure 3: Distribution of 7-grams in *Train* vs. *Target* set summaries (left) and vs. *BART* summaries (right).

4 Analysis of generated summaries

We perform lexical analysis of input abstracts, system generated summaries, and target summaries in our dataset, summarizing our findings below.

Input copying and synthesis To assess similarity between inputs and summaries, we first apply the evidence inference pipeline (Lehman et al., 2019; DeYoung et al., 2020)[6] to identify an evidence statement in each input document and classify it with an effect direction. Between each input evidence statement and the target and generated summaries, we compute ROUGE-1 scores. We compute the *Synthesis* rate as how often the effect direction agrees between the most similar evidence statement (by ROUGE-1 score) and the generated summary. In Table 3, we find that system generations match the effect of the closest input at a high rate (0.41-0.46), though no more frequently than we would expect based on the synthesis rate for the target summaries (0.48). Using ROUGE-1 scores, we also determine how often a generated summary is closer to an input document than the target (*Input Match*), which might indicate whether a system is performing an implicit synthesis by selecting an input and copying it. We find that systems sometimes copy inputs, but not in any consistent way.

***n*-gram self-repetition** Previously, Salkar et al. (2022) noted that models fine-tuned on the Cochrane corpus tend to generate summaries containing repeating patterns; however, they claim that the amount of such *self-repetition*[7] is fairly consistent between model-generated and human-written text. We analyze self-repetition rates for long n-grams (5- to 10-grams) and show that their occurrence rates are much higher in generated summaries than in human-written summaries. These long n-grams do not just represent stylistic patterns, but can contain important information such as the effect direction, e.g., "there is insufficient evidence to support the use" (see App. B for details), so the high rate of self-repetition is very concerning.

We find a clear distinction between generated and target summaries in the self-repetition of longer sequences, such as 7- to 10-grams (Figure 5 in App. B). Though the amount of self-repeating 10-grams in human-written summaries is negligible, it reaches over 80% in some of the examined models' outputs. The self-repetition rate for specific n-grams (the number of documents in which an n-gram appears) in generated summaries is also much higher than in the targets: some 7-grams occur in up to 70% of generated summaries (Figure 2; trends for other long n-grams are in App. B).

To determine the origin of these long n-grams, we calculate their overlap with summaries in the *Train* set and their corresponding input documents. While overlap with inputs is nearly zero, up to 90% of long n-grams are also found in *Train* set summaries (Figure 6 in App. C). Interestingly, models with global attention (LED or PRIMERA-based) seem to replicate more long sequences from the *Train* set summaries than BART-based ones, while in the Pegasus-based system (SciSpace) a smaller amount of self-repetition can be explained by fine-

[6] https://github.com/bwallace/RRnlp

[7] Salkar et al. (2022) define *self-repetition* as the proportion of generated summaries containing at least one n-gram of length $\geq$4 which also occurs in at least one other summary.

tuning. Finally, we observe that though the distributions of self-repeating n-grams in the target summaries of the *Test* set and *Train* set are very similar (Figure 3; left), in generated summaries the rate of self-repetition increases up to 500x compared to occurrence in the *Train* set summaries (Figure 3; right). Models amplify repeating patterns from the *Train* set to unnatural proportions!

5 Automated evaluation metrics

We compute automated metrics for each generated summary and include instance-level scores in our dataset. We investigate how these metrics correlate with other metrics (§5.1) and with human evaluation facets (§5.2).

Metrics from the MSLR leaderboard:

ROUGE: The leaderboard reported system-level ROUGE-1, ROUGE-2, and ROUGE-L F-scores (Lin, 2004). We report these same three metrics; in some plots, due to space constraints, we show the average of these three ROUGE metrics, which we call Avg-ROUGE-F.

BERTScore: We compute and report BERTScore-F (Zhang et al., 2020a) for each generated summary as computed using the RoBERTa-large model.

Delta-EI: We compute Delta-EI as introduced by Wallace et al. (2020) and modified by DeYoung et al. (2021) for the MSLR shared task. The metric computes the probability distributions of evidence direction for all intervention-outcome (I/O) pairs between inputs and the target and generated summaries. The final score is a sum over the Jensen-Shannon Divergence of probability distributions over all I/O pairs. Lower values indicate higher similarity to the target summary.

Other metrics we propose and examine:

NLI/STS/ClaimVer: These metrics leverage Sentence-BERT (Reimers and Gurevych, 2019) and are computed as the cosine similarity between the embedding of the target summary and the embedding of the generated summary when encoded with trained SBERT models. We use three pretrained variants of SBERT: RoBERTa fine-tuned on SNLI and MultiNLI (NLI); RoBERTa fine-tuned on SNLI, MultiNLI, and the STS Benchmark (STS); and PubMedBERT fine-tuned on MS-MARCO and the SciFact claim verification dataset (ClaimVer).

PIO-Overlap: Following Otmakhova et al. (2022a), we employ a strong PIO extractor (BioLinkBERT (Yasunaga et al., 2022) trained on EBM-NLP (Nye et al., 2018)) to extract PIO spans. For each target-generated pair, we define PIO-Overlap as the intersection of the two extracted sets of PIO spans normalized by the number of PIO spans in the target summary. Spans are only considered to overlap if they have the same label and one span is a subspan of the other.

Metric	Flu.	PIO	Dir.	Str.
ROUGE	-0.014	-0.010	0.007	-0.035
BERTScore	-0.000	0.022	0.036	-0.033
Delta-EI	0.066	-0.080	-0.060	-0.054
ClaimVer	-0.051	0.142**	-0.017	-0.093*
NLI	-0.026	0.053	-0.011	-0.063
STS	-0.042	0.066	0.001	-0.056
PIO-Overlap	0.043	0.358**	0.033	0.050

Table 4: Correlation coefficients between automated metrics and human evaluation facets. There is weak to no correlation between metrics and human-assessed facets (aside from between PIO-overlap and PIO). Statistical significance at $\alpha = 0.05$ is marked with *, 0.01 with **, though these thresholds for significance do not account for multiple hypothesis testing.

5.1 Correlation between automated metrics

We compute Pearson's correlation coefficients between pairs of metrics (Figure 8 in App. E). Most automated metrics are significantly correlated ($p < 0.01$), except Delta-EI and PIO-Overlap. ROUGE and BERTScore show a strong positive correlation ($r = 0.75$), and NLI and STS have a strong positive correlation ($r = 0.92$), unsurprising since the underlying models are trained on similar data. Delta-EI presents as bimodal, with two peaks around 0 and 1. Distributions of instance-level automated metrics per system are shown in App. D.

System ranks (§6) produced by automated metrics are highly correlated except for PIO-Overlap, which is anti-correlated (Figure 1). Ordering systems based on these metrics generally result in the same or similar rankings ($\rho \geq 0.77$ for all pairs of metrics besides PIO-Overlap), e.g., rankings from ClaimVer, NLI, and STS are identical ($\rho = 1$).

5.2 Correlation between automated metrics and human judgements

We investigate the relationship between automated metrics and human facet-based annotations. For this analysis, we normalize human facets to 4 agreement scores: Fluency, PIO, Direction, and Strength, each in the range [0, 1] (details in App. F).

Correlation coefficients between automated metrics and these four agreement scores are given in Table 4; PIO correlations are plotted in Figure 10 in App E. In general, there is weak to no correlation between metrics and human-assessed Fluency, PIO, Direction, and Strength, suggesting that automated metrics may not be adequately capturing aspects of summaries that humans determine to be important. The exception is PIO-Overlap, which has a statistically significant correlation with human-assessed PIO agreement, and presents as a promising future metric for the MSLR task; ClaimVer is also weakly correlated with PIO agreement.

Disappointingly, Delta-EI does not correlate with human-assessed Direction agreement. We investigate this further by computing empirical cumulative distribution functions (ECDFs) for each of the metrics w.r.t. Direction agreement (App. E). Delta-EI exhibits a small but desirable difference between instances where Direction agrees and instances where Direction disagrees (Agrees is more likely to have lower Delta-EI scores than Disagrees). In sum, Delta-EI shows some promise in detecting differences in Direction agreement, though further refinement of the metric is needed.

6 Comparing system rankings

Evaluation metrics for summarization can be used in two settings, to judge performance at the *instance* level (comparing individual summaries) or at the *system* level (comparing model performance over many instances). Here, we compare system-level rankings produced by automated metrics, human facet evaluation, and pairwise preference annotations to determine whether automated metrics effectively rank systems as humans would.

System rankings are computed by averaging the instance-level metric values or scores across all review instances for each system, and ranking from best to worst average score (direction depends on metric; higher is better for all scores except Delta-EI). We only average metrics over the subset of reviews for which we have human annotations. This ensures a fair comparison in the circumstance where we have selected an annotation sample that a system performs particularly well or poorly on. By doing this, the system rankings we present here are different than those computed using the same metrics from the MSLR leaderboards. We do not intend our computed rankings to be interpreted as the true system ranking; our analysis focuses on whether automated metrics and human evaluation are able to produce *similar* rankings of systems. Table 1 shows rankings as assessed by all automated metrics and human scores; Figure 1 shows Spearman correlation coefficients.

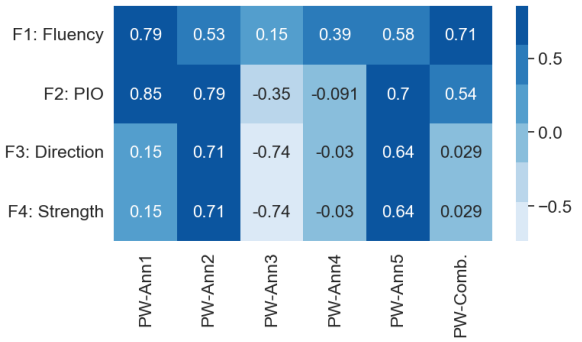

Figure 4: Spearman rank correlations between system ranks for each pairwise annotator and ranks derived from facet-based annotation. Annotators weigh quality facets differently when performing pairwise judgments.

Rankings by automated metrics are not correlated with rankings by human evaluation In general, system rankings from commonly reported automated metrics are not correlated or anti-correlated (lighter blue) with system rankings produced by human judgments. System rankings from automated metrics are highly correlated among themselves (ρ close to 1), aside from PIO-Overlap. PIO-Overlap rankings are strongly correlated with rankings from human PIO agreement. PIO-Overlap and Delta-EI ranks also correlate with the combined pairwise rankings, again suggesting that these two metrics may be the most promising for capturing human notions of summary quality.

Pairwise assessments do not weigh facets equally Pairwise-combined rankings are correlated with facet-based rankings for Fluency and PIO, but not Direction and Strength of claim. This may indicate that Fluency and PIO are more detectable problems, or that issues in Fluency and PIO are more prevalent in our data. The rank correlations also show that Direction and Strength are highly correlated and may capture similar aspects of system-level summary quality, making the case for dropping one of the two (likely Strength) in future annotations.

Pairwise preferences suggest that annotators weigh facets differently In Figure 4, we show Spearman correlation coefficients of facet-based rankings against the rankings of five pairwise annotators and the combined pairwise ranking. These

coefficients suggest that annotators weigh facets differently when comparing system output. Annotator 1 ranks similarly to Fluency and PIO facets, Annotators 2 and 5 rank similarly to PIO and Direction facets, while Annotators 3 and 4's rankings are uncorrelated with most facets.

7 Related work

Beyond ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020a), an extensive list of n-gram (Papineni et al., 2002; Banerjee and Lavie, 2005) and model-based (Zhao et al., 2019; Gao et al., 2020; Martins et al., 2020; Sellam et al., 2020; Yuan et al., 2021) summarization evaluation metrics have been proposed in the literature. In particular, model-based approaches that use question generation and question answering (Wang et al., 2020; Durmus et al., 2020; Deutsch et al., 2021) or NLI-based models (Kryscinski et al., 2020) have been proposed to assess summary factual consistency. Fabbri et al. (2021) and Deutsch et al. (2022) provide more thorough evaluations of many of these metrics on select summarization tasks. We perform evaluations using metrics previously reported on the MSLR task, and leave a systematic evaluation of metrics on this task and others to future work.

In Zhang et al. (2020b), the authors performed fact verification on generated radiology reports using an information extraction module, by aligning the extracted entities with entities found in the reference summary. Our PIO-Overlap metric similarly uses a PIO entity extraction module to assess concept overlap between generated and reference summaries. Falke et al. (2019) proposed to use NLI models to rank summaries by average entailment score per sentence against the input documents; this shares similarities with the Delta-EI score we evaluated, which attempts to quantify agreement relative to the reference summary with respect to the direction of evidence reported.

Deutsch et al. (2022) investigated system-level rankings produced by automated metrics and human evaluation and found minimal correlation between them, a finding corroborated by our work. Liu et al. (2022) introduced the robust summarization evaluation (RoSE) benchmark, containing human judgments for system outputs on the CNN/DM, XSum, and SamSum datasets. We extend such work into a novel domain (medical MDS for literature review) and demonstrate differences in automated metric performance and human evaluation in our domain and task. For example, though ROUGE correlates with human preferences in single-document (CNN/DM) and multi-document (MultiNews) news summarization, we find that it is poorly correlated with human judgments and preferences in the MSLR task.

Recent developments in large language modeling have also shifted the goalposts for evaluation. Goyal et al. (2022) found that although humans overwhelmingly prefer zero-shot GPT-3 summaries for news summarization, automated metrics were unable to capture this preference; they introduced a benchmark of human judgments and rationales comparing system outputs on the single-document news summarization task. More recently, Shaib et al. (2023) demonstrated that GPT-3 can be adapted for the MSLR task, and though the model outputs are generally found by human annotators to be faithful to the inputs, in the MDS setting the evidence direction often disagrees with the reference. Detecting these disagreements and developing automated metrics that can capture such disagreements are valuable pursuits and one of the motivations for our work. Further investigation into whether automated metrics developed using limited human evaluation benchmarks such as the dataset we introduce here will be a goal for future work.

8 Discussion

MDS for literature review may involve notions of summary quality not readily captured by standard summarization evaluation metrics. For example, our lexical analysis of generated summaries reveals a concerning level of self-repetition behavior, which is not penalized by standard metrics. Through two independent human evaluations (facet-based and pairwise preferences), we also show that automated metrics such as ROUGE and BERTScore are poorly correlated or even anti-correlated with human-assessed quality. This is not to say that these metrics do not provide any utility. Rather, further work is needed to understand what aspects of summary quality these metrics capture, and how to use them in combination with other metrics, novel metrics yet unintroduced, as well as human evaluation to better assess progress. We note that ours is not a systematic analysis of all automated summarization evaluation metrics, but is a focused study on evaluation metrics reported for the MSLR shared task and which we introduce under the hypothesis that they may be useful for capturing some

quality facets associated with this task. For those interested in the former, please refer to studies such as Fabbri et al. (2021) or Deutsch et al. (2022).

A positive finding from our work is the promise of the PIO-Overlap and Delta-EI metrics. Delta-EI shows some potential to capture evidence directional agreement between summaries, though the metric as currently implemented is noisy and does not cleanly separate summaries that agree and disagree on direction. PIO-Overlap, a metric we introduce, correlates with human-assessed PIO agreement, suggesting that it could be a performant, scalable alternative to human evaluation of this quality facet. Still, more work is needed to probe how variants of these metrics could be adapted to evaluate performance on MSLR and other MDS tasks.

Finally, we note that human evaluation is difficult because people value different qualities in summaries. The rank-based analysis we perform does not account for interactions between related quality facets and is unable to elicit relationships between overall quality and individual quality facets. The majority of pairwise preference annotations in our dataset also include short free text justifications for preference decisions, which could be used to further study this problem. Other promising directions for future work involve studying how to optimally elicit human preferences, such as how to sample instances for labeling to maximize our confidence in the resulting system-level rankings.

9 Conclusions

There have been major recent advances in the generative capabilities of large language models. Models like ChatGPT,[8] GPT-3 (Brown et al., 2020), and PubmedGPT[9] demonstrate aptitude on many tasks but have also been shown to confidently produce factually incorrect outputs in specialized and technical domains.[10] Medicine is a specialized domain where incorrect information in generated outputs is difficult to identify and has the potential to do harm. There is therefore a pressing need for the community to develop better methods to assess the quality and suitability of generated medical texts. Our investigation confirms that there is significant room for improvement on medical MDS evaluation. We hope that the resources and findings we contribute in this work can assist the community towards this goal.

Limitations

Though we include 6 systems in our annotation which reflect the current state-of-the-art, all of the models are Transformer-based and fine-tuned on just the Cochrane dataset, which may limit the diversity of our generated summaries. Additionally, none of the systems are generating summaries that approach the accuracy of human-written summaries. As a consequence, though the summaries in our dataset span the spectrum of quality, they may have less coverage on the higher end of quality (summaries approaching the accuracy and utility of human-written review summaries).

Our analysis of evaluation metrics also assumes the existence of reference summaries. In many real-world summarization scenarios, reference summaries do not exist, and reference-free evaluation metrics are needed for assessment. We refer the reader to related work in reference-free summarization evaluation (Vasilyev et al., 2020; Gao et al., 2020; Luo et al., 2022), which have been found in some settings by Fabbri et al. (2021) to exhibit even lower correlation with human notions of summary quality; the performance of these metrics on MSLR evaluation is unknown and is left to future work.

Our notions of summary quality also do not necessarily correspond to clinical utility. As with anything in the medical setting, it is of utmost importance to verify correctness and the quality of evidence before using any generated text to make or guide clinical decisions.

Ethical Considerations

As with other applications of NLP in the medical domain, results of MSLR systems must be verified by domain experts before they should be considered for use in clinical guidance. We do not intend the system outputs included in our dataset and analysis to be used for such end applications, as this would be clearly premature given the low quality of generated summaries and our lack of ability to assess the prevalence of factuality errors in these summary texts. Nonetheless, we believe that medical MDS holds eventual promise, and it is of vital importance that we study its challenges and how to measure and detect quality issues in generated text.

[8] https://openai.com/blog/chatgpt

[9] https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b

[10] Stack Overflow banned ChatGPT responses due to the high rate of inaccurate and misleading information.

Acknowledgements

This research was partially supported by National Science Foundation (NSF) grant RI-2211954, by the National Institutes of Health (NIH) under the National Library of Medicine (NLM) grant 2R01LM012086. YO and THT are supported by the Australian Government through the Australian Research Council Training Centre in Cognitive Computing for Medical Technologies (project number ICI70200030).

References

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

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

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.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. 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. 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.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MS2: Multi-document summarization of medical studies. In *EMNLP*.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, 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.

Peter Emerson. 2013. The original borda count and partial voting. *Social Choice and Welfare*, 40:353–358.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

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

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. 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.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2022a. Exploring the challenges of open domain multi-document summarization. *ArXiv*, abs/2212.10526.

John Giorgi et al. 2022b. MSLR leaderboard: led-base-16384-ms2. https://leaderboard.allenai.org/mslr-ms2/submission/ccfknkbml1mljnftf7d0. Accessed: 2022-09-15.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *ArXiv*, abs/2209.12356.

Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of pico as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*, volume 2006, page 359. American Medical Informatics Association.

John PA Ioannidis. 2016. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, 94(3):485–514.

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.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. *arXiv preprint arXiv:1904.01606*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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.

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

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

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. 2022. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *ArXiv*, abs/2212.07981.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Ge Luo, Hebi Li, Youbiao He, and Forrest Sheng Bao. 2022. PrefScore: Pairwise preference learning for reference-free summarization quality assessment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5896–5903, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2020. Sparse text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4252–4273, Online. Association for Computational Linguistics.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Yulia Otmakhova, Thinh Hung Truong, Timothy Baldwin, Trevor Cohn, Karin Verspoor, and Jey Han Lau. 2022a. LED down the rabbit hole: exploring the potential of global attention for biomedical multi-document summarisation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 181–187, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022b. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111, Dublin, Ireland. Association for Computational Linguistics.

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.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 4768–4779, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nikita Salkar, Thomas Trikalinos, Byron Wallace, and Ani Nenkova. 2022. Self-repetition in abstractive neural summarizers. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 341–350, Online only. Association for Computational Linguistics.

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

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain James Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success).

Zejiang Shen, Kyle Lo, Lauren Jane Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *ArXiv*, abs/2206.10883.

Kartik Shinde, Trinita Roy, and Tirthankar Ghosal. 2022. An extractive-abstractive approach for multi-document summarization of scientific articles for literature review. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Rahul Tangsali, Aditya Vyawahare, Aditya Mandke, Onkar Litake, and Dipali Kadam. 2022. Abstractive approaches to multidocument summarization of medical literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. Scifact-open: Towards open-domain scientific claim verification. *ArXiv*, abs/2210.13777.

Byron C. Wallace, Sayantani Saha, Frank Soboczenski, and Iain James Marshall. 2020. Generating (factual?) narrative summaries of RCTs: Experiments with neural multi-document summarization. In *AMIA Annual Symposium*.

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.

Ruben Wolhandler, Arie Cattan, Ori Ernst, and Ido Dagan. 2022. How "multi" is multi-document summarization? *ArXiv*, abs/2210.12688.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *ArXiv*, abs/2106.11520.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *ArXiv*, abs/2007.14062.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. *ArXiv*, abs/1904.09675.

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 *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics, pages 5108–5120, 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.

A Facet-based Annotation

The questions and answer options shown to annotators for facet annotation are shown in Table 5. If merging all Yes and Partial Yes classes, agreement proportion between annotators increases for Fluency (0.87 → 0.97), Population (0.56 → 0.64), Intervention (0.77 → 0.90), and Outcome agreement (0.36 → 0.44).

B Self-repetition rates in generated summaries

Most of the long n-grams repeating across documents contain meaningful statements regarding the direction or strength of effect findings rather than purely stylistic patterns, which means that the systems are prone to introducing factuality mistakes by replicating common statements. In Table 6 we show the examples of the most repetitive 8-grams for the 6 models, together with the percentage of generated summaries they occur in.

We also show that the self-repetition rate for n-grams with n > 4 have very dissimilar trends for generated summaries in comparison to human-written summaries (Figure 5) The amount of 5-grams and higher self-repetition also differs between models .

C Copying self-repeating n-grams from training set

In Figure 6, we show the percentages of self-repeating n-grams from generated summaries which can also be found in the target summaries in the Train set.

D Automated metric distributions per system

Distributions of automated metrics for all instances per system are shown in Figure 7.

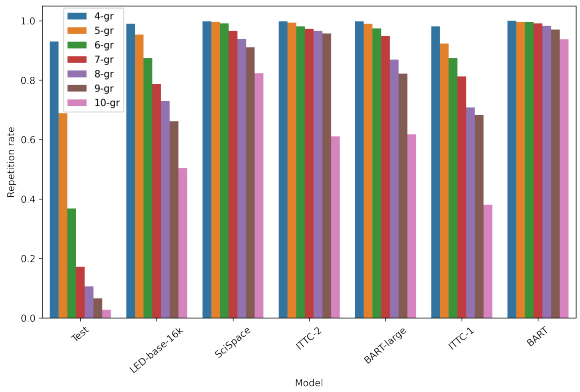

Figure 5: Rate of self-repetition for models generations and the human written summaries (*Test*)

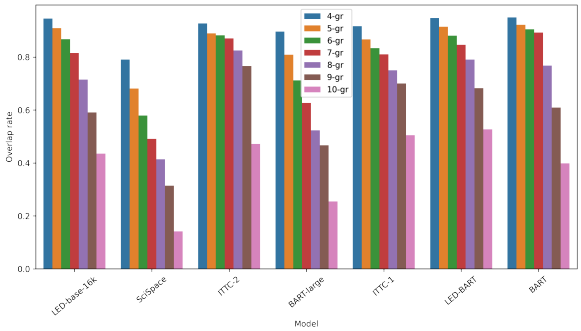

Figure 6: Percentage of self-repeating n-grams in generated summaries which also occur in the target summaries of the Train set.

E Correlations between metrics in the Cochrane dataset

We present correlations between all automated metrics along with correlation coefficients (Figure 8). ROUGE and BERTScore are strongly correlated. NLI and STS are strongly correlated. Delta-EI has a bimodal distribution. PIO-Overlap is uncorrelated with other metrics.

Correlations between automated metrics and the normalized PIO facet score are shown in Figure 9. In general, automated metrics are poor predictors of PIO agreement, except PIO-Overlap, which is positively correlated with PIO agreement ($p < 0.05$). This confirms that model extraction and alignment of PIO spans is a promising direction for assessing PIO agreement. ClaimVer also shows a weak but statistically significant correlation with PIO agreement. The ClaimVer metric is computed based on embedding similarity between two texts using a model trained on the SciFact scientific claim verification dataset (Wadden et al., 2020); the SciFact task measures whether evidence entails or refutes

Question	Answer options
1. Is the generated summary fluent?	2: Yes–there are no errors that impact comprehension of the summary 1: Somewhat, there are some minor grammatical or lexical errors, but I can mostly understand 0: No, there are major grammatical or lexical errors that impact comprehension
2. Is the *population* discussed in the generated summary the same as the population discussed in the target summary?	2: Yes 1: Partially 0: No N/A: No population in generated summary Other: Comment
3. Is the *intervention* discussed in the generated summary the same as the intervention discussed in the target summary?	2: Yes 1: Partially 0: No N/A: No intervention in generated summary Other: Comment
4. Is the *outcome* discussed in the generated summary the same as the outcome discussed in the target summary?	2: Yes 1: Partially 0: No N/A: No outcome in generated summary Other: Comment
5. What is the effect direction in the *target* summary for the main intervention and outcome considered?	(+1): Positive effect 0: No effect (-1): Negative effect N/A: no effect direction is specified in the target summary Other: Comment
6. What is the effect direction in the *generated* summary for the main intervention and outcome considered?	(+1): Positive effect 0: No effect (-1): Negative effect N/A: no effect direction is specified in the generated summary Other: Comment
7. What is the strength of the claim made in the *target* summary?	3: Strong claim 2: Moderate claim 1: Weak claim 0: Not enough evidence (there is insufficient evidence to draw a conclusion) N/A: No claim (there is no claim in the summary) Other: Comment
8. What is the strength of the claim made in the *generated* summary?	3: Strong claim 2: Moderate claim 1: Weak claim 0: Not enough evidence (there is insufficient evidence to draw a conclusion) N/A: No claim (there is no claim in the summary) Other: Comment

Table 5: Questions and answer options used during facet annotation.

Model	Most frequent 8-gram	Self-repetition rate (%)
Targets	*the conclusions of the review once assessed .*	1.5
LED-base-16k	*there is insufficient evidence to support or refute*	9.4
ITTC-1	*there is insufficient evidence to support the use*	18.7
BART-large	*there is insufficient evidence to support the use*	22.8
SciSpace	*there is insufficient evidence to support the use*	55.5
BART (baseline)	*there is insufficient evidence from randomised controlled trials*	59.1
ITTC-2	*there is insufficient evidence to support the use*	65.1

Table 6: Examples of 8-grams which are most frequently repeated across generated summaries, together with their self-repetition rate.

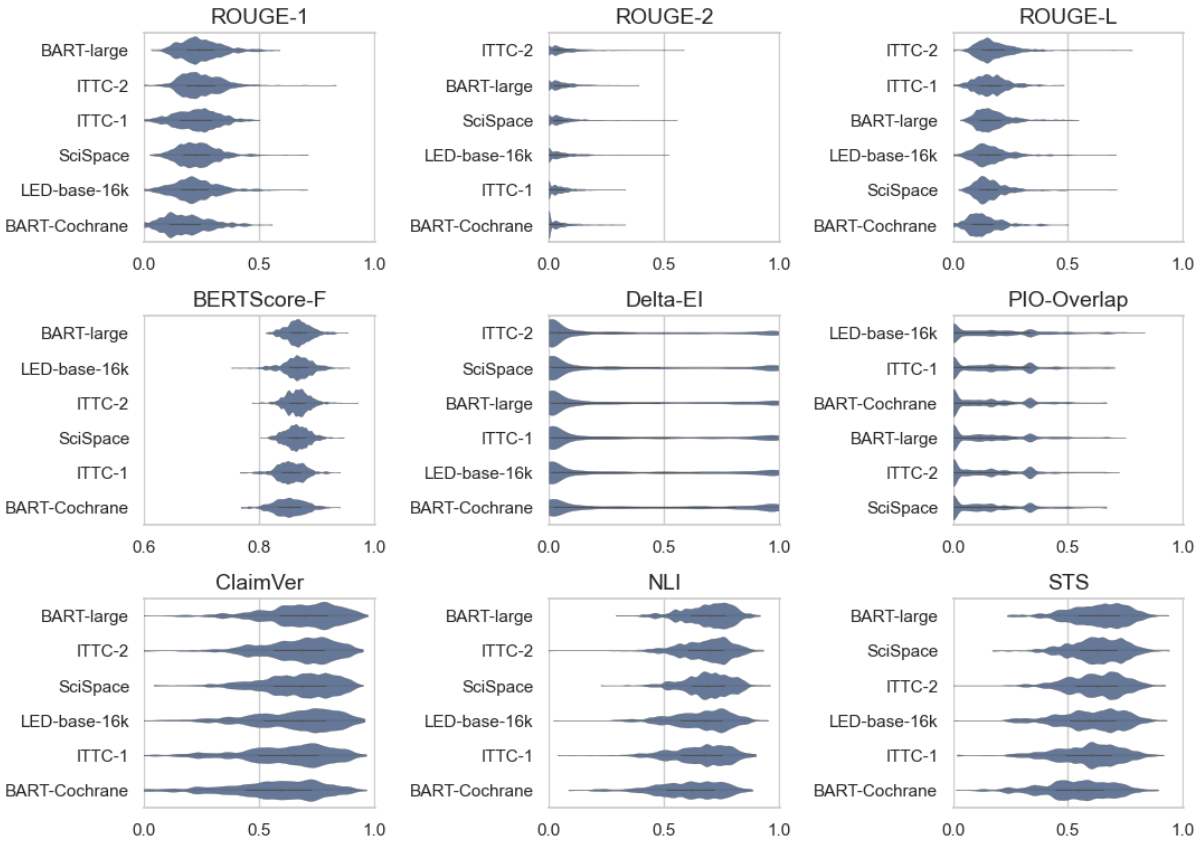

Figure 7: Distribution of instance-level automated metric values by system (n=470 for each system). Each subplot is sorted on system from best to worst by median score. The median score ranking is typically not identical to the ranking of the corresponding metric from the MSLR leaderboard, which are computed based on micro-averaged metric values.

a scientific claim, which is somewhat analogous to our evaluation task for medical multi-document summarization.

We also assess whether metrics can distinguish between summaries where the Direction agrees with the target and summaries where the Direction disagrees. We present the empirical cumulative distribution functions (ECDF) for each automated metric, showing the separation of metrics between when Direction agrees and disagrees (Figure 10. The Delta-EI metric is somewhat sensitive to human-assessed directional agreement (a higher proportion of generated summaries where the Direction agrees with the target have lower Delta-EI scores), though we note that the difference is small. PIO-Overlap also shows some separation between the two Direction classes (a higher proportion of disagrees have lower PIO-Overlap score than agrees), though again the difference is subtle.

F Normalizing human facet scores

Responses to the Fluency question result in a 3-class ordinal variable that we map to the range [0, 1], where 0.0 is disfluent, 0.5 is somewhat fluent,

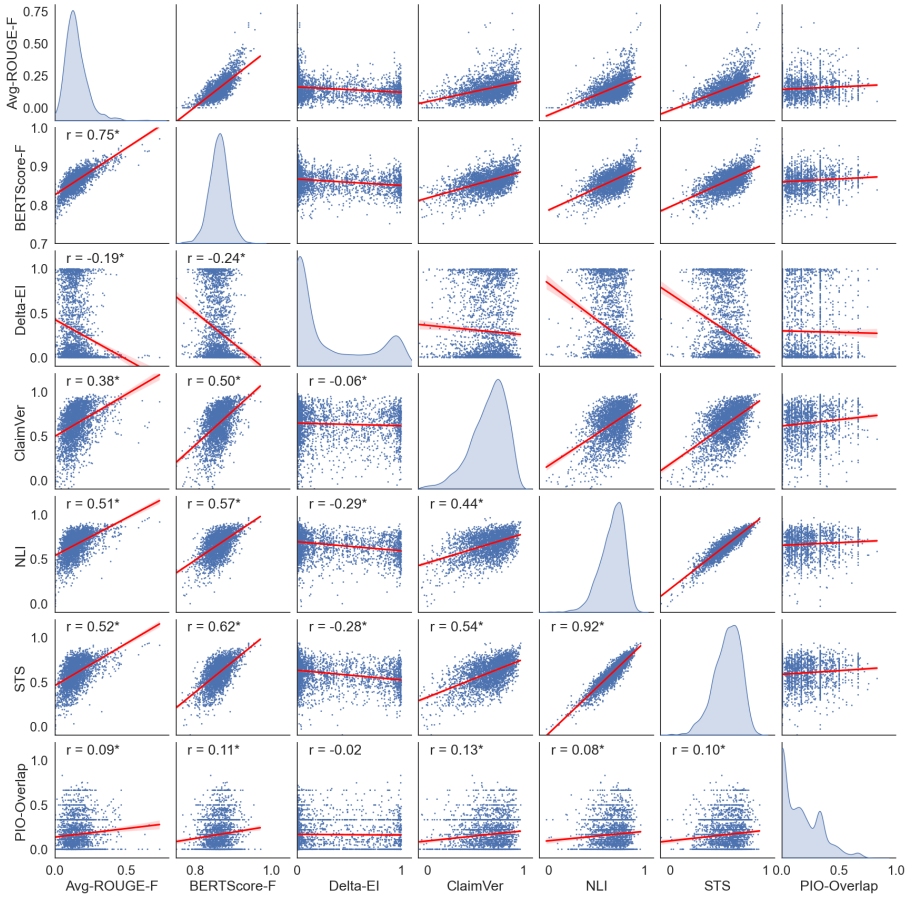

Figure 8: Correlations between automated metrics in the Cochrane dataset. Pearson's correlation coefficients (r) are shown, along with an * if $p < 0.01$.

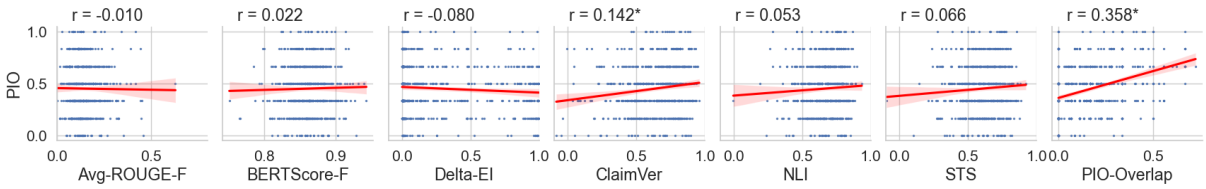

Figure 9: Correlations between automated metrics and the normalized PIO facet score.

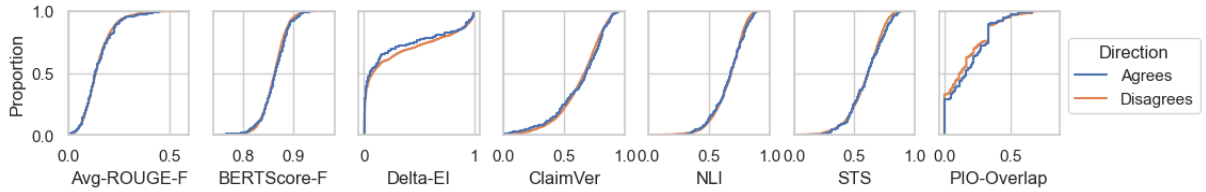

Figure 10: Empirical cumulative distribution function (ECDF) of each of the automated metrics and their values for summaries where humans assessed the evidence direction to Agree versus those assessed to Disagree.

and 1.0 is fluent. PIO aggregates agreement over Population, Intervention, and Outcome, where each of P, I, and O are 3-class ordinal variables that we map to the range [0, 1] as we do Fluency; we average the three facets to get PIO agreement. For evidence direction, though each of the two annotated questions has 4 answers (positive, no effect, negative, or no direction given), we elect to define Direction as a binary class. We normalize Direction to 1 if the target direction and generated direction agree and 0 if they disagree. For Strength, each of the two annotated questions has 4 answers (strong,

moderate, weak, and not enough evidence). We take the difference between the answers for the target and generated summaries and normalize to the range [0, 1] to yield our Strength agreement score.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section, and parts of 8. Discussion

☑ A2. Did you discuss any potential risks of your work?
Ethical considerations section, and parts of 8. Discussion

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

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

B ☑ Did you use or create scientific artifacts?

3. Dataset; we create a dataset in this work and describe how we go about collecting data

☑ B1. Did you cite the creators of artifacts you used?
Throughout the paper

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Licensing for our data artifact will be available on Github

☑ 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)?
Throughout the paper, also in 3. Dataset and 8. Discussion

☐ 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?
Not applicable. No names and unique identifiers are included in the dataset

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We discuss the provenance of data in our dataset in 2. Background and 3. Dataset

☑ 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.
All reported in 3. Dataset

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

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?
No response.

☐ 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?
No response.

☐ 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.)?
No response.

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

Section 3, 5, and 6 discuss our human annotation protocols

☑ 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.?
Full text is in Appendix; a brief description in Section 3

☐ 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)?
Not applicable. Annotators are included as authors on the paper

☐ 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?
Not applicable. Annotators are included as authors on the paper

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3 describes annotator demographics and background