Shortcomings of Question Answering Based Factuality Frameworks for Error Localization

Ryo Kamoi **Tanya Goyal** **Greg Durrett**
Department of Computer Science
The University of Texas at Austin
ryokamoi@utexas.edu

Abstract

Despite recent progress in abstractive summarization, models often generate summaries with factual errors. Numerous approaches to detect these errors have been proposed, the most popular of which are question answering (QA)-based factuality metrics. These have been shown to work well at predicting summary-level factuality and have potential to localize errors within summaries, but this latter capability has not been systematically evaluated in past research. In this paper, we conduct the first such analysis and find that, contrary to our expectations, QA-based frameworks fail to correctly identify error spans in generated summaries and are outperformed by trivial exact match baselines. Our analysis reveals a major reason for such poor localization: questions generated by the QG module often inherit errors from non-factual summaries which are then propagated further into downstream modules. Moreover, even human-in-the-loop question generation cannot easily offset these problems. Our experiments conclusively show that there exist fundamental issues with localization using the QA framework which cannot be fixed solely by stronger QA and QG models.

1 Introduction

Although abstractive summarization systems (Rush et al., 2015; See et al., 2017; Lewis et al., 2020) have improved drastically over the past few years, these systems often introduce factual errors into generated summaries (Cao et al., 2018; Kryscinski et al., 2019). Recent work has proposed a number of approaches to detect these errors, including using off-the-shelf entailment models (Falke et al., 2019; Laban et al., 2022), question answering (QA) models (Chen et al., 2018; Wang et al., 2020; Durmus et al., 2020), and discriminators trained on synthetic data (Kryscinski et al., 2020). Such methods have also been explored to identify error spans within summaries (Goyal and Durrett, 2020)

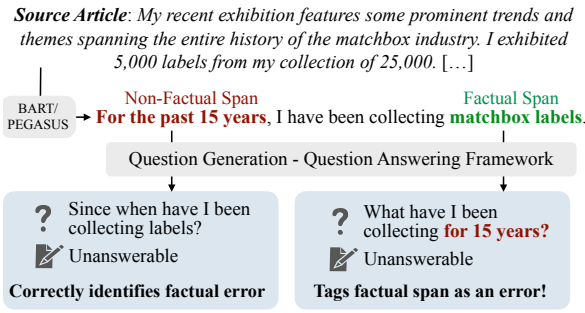


Figure 1: Factual error localization using QA metrics. Questions are generated for summary spans and then answered by a QA model using the source article as context. For factual spans (e.g. *matchbox labels*), we expect the predicted answers to match the original spans. However, non-factual spans in generated questions inherited from summaries may render these unanswerable and lead to incorrect error localization.

and perform post-hoc error correction (Dong et al., 2020; Cao et al., 2020).

Among these different approaches for evaluating factuality, QA-based frameworks are the most widely adopted (Chen et al., 2018; Scialom et al., 2019; Durmus et al., 2020; Wang et al., 2020; Scialom et al., 2021; Fabbri et al., 2022). These evaluate the factuality of a set of spans in isolation, then combine them to render a summary-level judgment. Figure 1 illustrates the core mechanism: question generation (QG) is used to generate questions for a collection of summary spans, typically noun phrases or entities, which are then compared with those questions' answers based on the source document to determine factuality. Due to this span-level decomposition of factuality, QA frameworks are widely believed to localize errors (Chen et al., 2018; Wang et al., 2020; Gunasekara et al., 2021). Therefore, the metrics have been applied in settings like post-hoc error correction (Dong et al., 2020), salient (Deutsch and Roth, 2021) and incorrect (Scialom et al., 2021) span detection, and text alignment (Weiss et al., 2021). However, their ac-

tual span-level error localization performance has not been systematically evaluated in prior work.

In this paper, we aim to answer the following question: **does the actual behavior of QA-based metrics align with their motivation?** Specifically, we evaluate whether these models successfully identify error spans in generated summaries, independent of their final summary-level judgment. We conduct our analysis on two recent factuality datasets (Cao and Wang, 2021; Goyal and Durrett, 2021) derived from pre-trained summarization models on two popular benchmark datasets: CNN/DM (Hermann et al., 2015; Nallapati et al., 2016) and XSum (Narayan et al., 2018). Our results are surprising: **we find that good summary-level performance is rarely accompanied by correct span-level error detection.** Moreover, even trivial exact match baselines outperform QA metrics at error localization. Our results clearly show that although motivated by span-level decomposition of the factuality problem, the actual span-level predictions of QA metrics are very poor.

Next, we analyze these failure cases to understand why QA-based metrics diverge from their intended behavior. We find that the most serious problem lies in the question generation (QG) stage: generated questions for non-factual summaries inherit errors from the input summaries (see Figure 1). This results in poor localization wherein factual spans get classified as non-factual due to presupposition failures during QA. Furthermore, we show that such inherited errors cannot be easily avoided: decreasing the length of generated questions reduces the number of inherited errors, but very short questions can be under-specified and not provide enough context for the QA model. In fact, replacing automatic QG with human QG also does not improve the error localization of QA metrics. These results demonstrate fundamental issues with the current QA-based factuality frameworks that cannot be patched by stronger QA/QG methods.

Our contributions are as follows. (1) We show that QA-based factuality models for summarization exhibit poor error localization capabilities. (2) We provide a detailed study of factors in QG that hamper these models: inherited errors in long generated questions and trade-offs between these and short under-specified questions. (3) We conduct a human study to illustrate the issues with the QA-based factuality framework independent of particular QA or QG systems.

2 QA-Based Factuality Metrics

Recent work has proposed numerous QA-based metrics for summarization evaluation, particularly factuality (Chen et al., 2018; Scialom et al., 2019; Eyal et al., 2019; Durmus et al., 2020; Wang et al., 2020; Deutsch and Roth, 2021). These proposed metrics follow the same basic framework (described in Section 2.1), and primarily differ in the choice of off-the-shelf models used for the different framework components (discussed in Section 2.2).

2.1 Basic Framework

Given a source document D and generated summary S, the QA-based metrics output a summary-level factuality score y_S that denotes the factual consistency of S. This includes the following steps (also outlined in Figure 2):

1. **Answer Selection**: First, candidate answer spans $a_i \in S$ are extracted. These correspond to the base set of *facts* that are compared against the source document D. Metrics evaluated in this work (Scialom et al., 2021; Fabbri et al., 2022) consider all noun phrases and named entities in generated summaries as the answer candidates set, denoted by $span(S)$.
2. **Question Generation**: Next, a question generation model (G) is used to generate questions for these answer candidates with the generated summary S as context. Let $q_i = G(a_i, S)$ denote the corresponding question for span a_i.
3. **Question Filtering**: Questions for which the question answering (A) model's predicted answer $A(q_i, S)$ from the summary does not match the original span a_i are discarded, i.e., when $a_i \neq A(q_i, S)$. This step is used to ensure that the effects of erroneous question generation do not percolate down the pipeline; however, answer spans that do not pass this phase cannot be evaluated by the method.
4. **Question Answering**: For each generated question q_i, the A model is used to predict answers using the source document D as context. Let $p_i = A(q_i, D)$ denote the predicted answer.
5. **Answer Comparison**: Finally, the predicted answer p_i is compared to the expected answer a_i to compute a similarity score $sim(p_i, a_i)$. The overall summary score y_S is computed by averaging over all span-level similarity scores:

$$y_S = \frac{1}{|span(S)|} \sum_{a_i \in span(S)} sim(A(q_i, D), a_i)$$

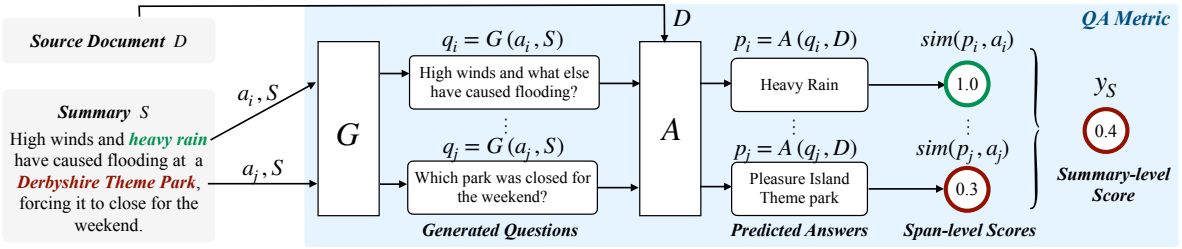


Figure 2: Overall workflow for the QA metrics. First, questions are generated for all NEs and NPs in the generated summary. Answers to these questions are obtained from the source document. Then, a factuality score is computed for each summary span based on it similarity with the predicted span from the previous step. Finally, all span-level scores are aggregated to obtain the final summary-level factuality.

Based on the motivation behind QA metrics, these similarity scores $sim(p_i, a_i)$ should indicate the factuality of the corresponding spans. If span a_i is factual, then the $G - A$ pipeline should output $p_i \in D$ with high similarity to a_i. Conversely, if a_i is non-factual, the similarity score $sim(p_i, a_i)$ should be low. While prior research has only evaluated their sentence-level performance, we use these span-level factuality scores to additionally evaluate the localization performance of QA metrics.

2.2 QA Metrics compared

In this work, we focus our analysis on the two best performing QA-based metrics from prior work:

QuestEval (QE) Scialom et al. (2021) generate questions for answer spans extracted from both the summary ("precision questions") and source document ("recall questions"). We only use the former in our experiments as these are shown to correlate better with factuality. Both the A and G components of QuestEval use T5-Large models (Raffel et al., 2020) fine-tuned on question answering datasets (Rajpurkar et al., 2018; Trischler et al., 2017). The similarity score $sim(p_i, a_i)$ in this framework is computed as the average of the lexical overlap, BERTScore, and the answerability score predicted by A.

QAFactEval (QAFE) Fabbri et al. (2022) conduct an ablation study over the different combinations of available A and G models. Here, we use their best-performing combination: an ELECTRA-based A model and a BART-based G model fine-tuned on the QA2D dataset (Demszky et al., 2018). The $sim(p_i, a_i)$ score is obtained using the learned metric LERC (Chen et al., 2020). If $A(q_i, D)$ is unanswerable for span a_i, QAFactEval sets the similarity score $sim(\_, a_i) = 0$ instead of using the LERC metric.

3 Experimental Setup

3.1 Task Definition

Given document D and a generated summary S, let $y^*_S \in \{0, 1\}$ denote the gold summary-level factuality label. Additionally, we assume access to $L = \{(a, y^*_a)\}$ which denotes the set of spans $a \in span(S)$ and their corresponding span-level gold factuality labels $y^*_a \in \{0, 1\}$.

First, we evaluate the **summary-level performance** of factuality models, i.e., is the predicted factuality equal to the gold factuality judgment y^*_S? To do this, we covert the predicted factuality score y_S to a binary judgment using dataset-specific thresholds. For each factuality model evaluated, we select thresholds that yield the best F1 scores on the validation set on each dataset.

Next, we evaluate the **span-level (localization) performance** of factuality models. Similar to the previous setting, we convert span-level predictions y_a to binary labels using the best-F1 threshold derived from the validation set. We report the macro-averaged performance at correctly predicting the span-level label $y^*_a \; \forall a \in span(S)$ across all (D, S) pairs in the evaluation dataset.

To align with the current QA frameworks, we restrict our evaluation to spans that correspond to named entities and noun phrases. This takes a generous view of the QA metrics' performance as it does not penalize them for failing to identify factual-errors outside NPs and NEs. This setting allows us to study the fundamental issues with the QA framework instead of those that can potentially be addressed by extending the question types considered in the framework.

Note that even for NP and NE spans, sometimes the QA metric does not return a span-level prediction if the span has not been selected as an answer candidate or has been discarded during the question filtering phase. We assume the predicted label $y_a = 1$ for such spans, as the model failed to detect any errors.[1] We discuss the performance loss due to this additional filtering step in Appendix C.

3.2 Datasets

We conduct our analysis on two human-annotated factuality datasets from prior work that provide gold annotations of factuality at the token level. To the best of our knowledge, these two are the only datasets that include span-level factuality annotation for summaries generated by SOTA models.

CLIFF (Cao and Wang, 2021) is a dataset consisting of summaries generated by BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020) models trained on the XSum and CNN/DM summarization datasets. For each generated summary, the dataset includes token-level factuality labels $y_t^* \in \{0, 1\}$. For $y_t^* = 0$, these are additionally labeled with fine-grained error types: extrinsic, intrinsic, or requiring world knowledge.

GD21 (Goyal and Durrett, 2021) contains XSum summaries generated using a fine-tuned BART model. Similar to CLIFF, it contains token-level factuality labels for all generated summaries.

Deriving gold summary- and span-level factuality labels from human annotations To derive the summary-level gold label y_S^* from these token-level human annotations, we set $y_S^* = 1$ iff all tokens are factual, i.e. $y_t^* = 1 \; \forall t \in S$. To derive span-level gold labels, for each NP/NE span a, we set $y_a^* = 1$ iff all tokens $t \in a$ are factual.

We construct validation and test sets by dividing each dataset into equal subsets. The statistics for the test set are included in Table 1. It shows that ~26% of non-factual tokens do not correspond to NEs or NPs and are therefore ignored by the QA metrics' evaluation pipeline. Also, note that the error statistics differ for the XSum summaries in GD21 and CLIFF due to the differences in the annotation methodologies and the trained models used (both BART and PEGASUS in CLIFF vs only BART in GD21).

[1]Operationally, we set $sim = 6.0$ for QAFactEval and $sim = 1.0$ for QuestEval for filtered spans.

Label Gran.	Metric	GD21 XSum	CLIFF C/D	CLIFF XSum
Summ.	Total	46	150	150
	% Non-Factual	52.2	15.3	70.7
Span	# per summary	7.9	15.3	5.4
	% Non-Factual	9.9	1.9	28.1
Token	# per summary	17.1	31.6	13.1
	% Non-Factual	8.8	1.7	24.5
	% Ignored (Non-Factual)	28.9	24.5	27.6

Table 1: Test set statistics for CLIFF and GD21 at different levels of label granularity. All our evaluation is done at the summary- and span-levels to align with the QA metrics' formulation. We convert the token-level human annotations to span-level to achieve this. The table reports the % of non-factual tokens outside NE/NPs that are ignored by the QA metrics' evaluation pipeline.

3.3 Baselines for Comparison

Exact Match Baseline (EM) first extracts all nouns, proper nouns, numbers, adjectives, and pronoun tokens from the generated summary S. For these tokens, we set $y_t = 1$ if $y_t \in D$, else $y_t = 0$. We use the fraction of tokens predicted as factual as a summary-level score.

Dependency-Arc Entailment (DAE) Goyal and Durrett (2020) evaluate the factuality of each dependency arc in generated summaries separately. We follow the methodology proposed by Goyal and Durrett (2021) to derive both summary- and token-level factuality scores from these arc-level judgments. We refer readers to the original paper for further details. We use their available model checkpoint in our experiments.[2]

We convert token-level judgments from these baseline models into span-level judgments to make their outputs compatible with our evaluation framework. This is described in detail in Appendix A.3.

4 Summary vs. Span Level Performance

QA metrics motivate the use of span-level factuality as building blocks for evaluating factuality at the sentence level. Therefore, **our hypothesis is that good summary-level performance must be accompanied by good span-level performance**. Here, we test this by comparing summary- and span-level performances.

Figure 3 outlines the performance of QA metrics and baseline systems. The top row shows ROC

[2]Code and trained model checkpoint provided by authors at: https://github.com/tagoyal/factuality-datasets

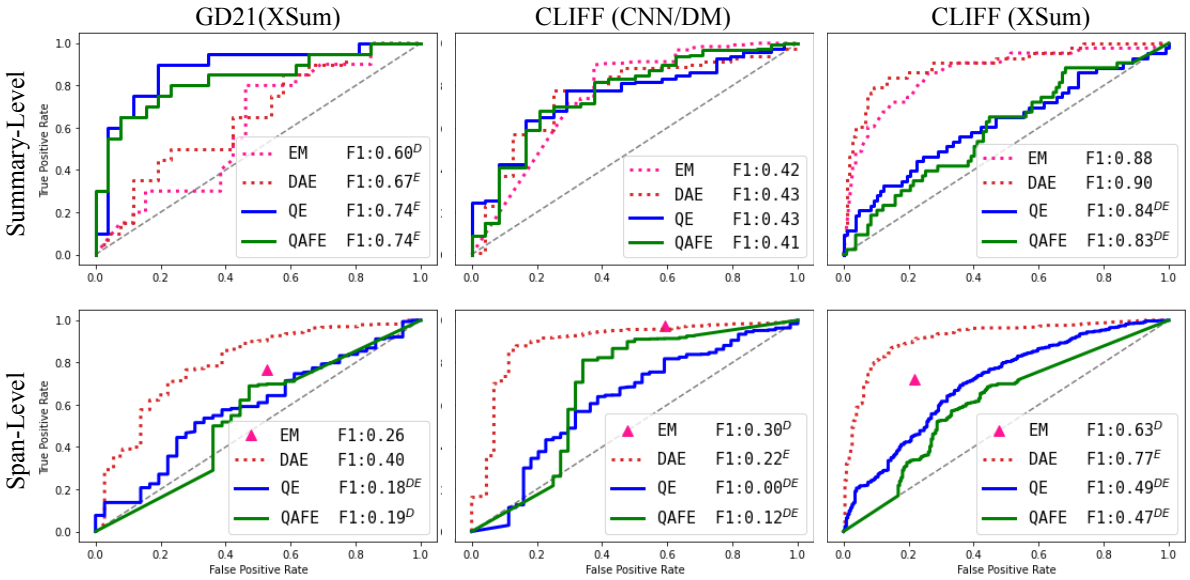


Figure 3: ROC curves and F1 scores (legend) for all systems at the summary- and span-levels (EM is a point as it provides hard binary judgments). Here, ^{D} (or ^{E}) denotes that the performance difference with DAE (or EM) is statistically significant according to a paired bootstrap test (p-value < 0.05). We observe that for QA metrics, good summary-level performance (e.g. on GD21 and CLIFF (CNN/DM)) does not imply good localization performance.

curves and F1 scores (in the legend) for all three datasets; the bottom row shows span-level results. The dotted black lines show the performance of a random baseline. First, we observe that none of the baselines or QA metrics have a clear advantage over other systems for all datasets at the summary-level. For instance, QA metrics outperform baselines on GD21, and show similar performance on CLIFF (CNN/DM) and worse performance on the CLIFF (XSum) dataset. However, **across all datasets, we see that there exists a substantial mismatch between the performance of QA metrics at the summary- and span-levels.** Notably, for GD21, both QE and QAFE substantially outperform baseline models at the summary-level, but exhibit much poorer span-level performance. Similarly, QA metrics are comparable to baselines at the summary-level for CLIFF (CNN/DM) but much worse at the span-level. On the other hand, the error localization performance of the DAE model is more consistent with its summary-level performance. Surprisingly, the trivial **exact match (EM) baseline consistently outperforms QA metrics at error localization for all datasets**. These results clearly show that our hypothesis is false: QA-based metrics do not provide reliable span-level explanations for their summary-level predictions.

Note that the diagonal lines in the span-level ROC curves for QA metrics arise due to a large number of spans being assigned the same factuality scores. As discussed in Section 2.1, some spans are filtered during the question filtering stage (Step 3) if their corresponding generated questions are of low quality. We consider these to be factual and assign them the maximum factuality score; this results in the diagonal line from $(0, 0)$.[3] We study the effects of this span filtering on localization performance in Appendix C. Additionally, QAFE assigns the same factuality score ($= 0$) to all spans with unanswerable questions resulting in the diagonal line to $(1, 1)$.

5 Why do QA metrics fail at span-level error localization?

Consider the error localization task in the example in Figure 1. Here, the QA metric needs to correctly distinguish between the factual span "*matchbox labels*" and the extrinsic error "*for the past 15 years*". For such summaries (containing a mix of factual and non-factual spans), we observed that the generated questions for factual spans often **inherit** non-factual summary spans. Given such questions, e.g. "*What have I been collecting for 15 years?*", an ideal QA model ***should*** predict unanswerable (even though that hurts localization) as the source article does not include any mention of an item being collected for 15 years. Based on this obser-

[3]Note that QE generates multiple questions for each span and therefore rarely discards spans (it is not likely that all questions are bad). Therefore, it has a shorter diagonal line.

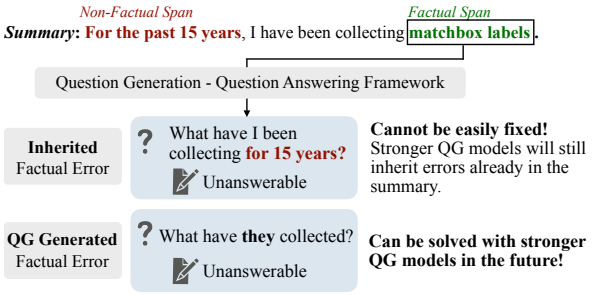


Figure 4: Generated questions broadly include two types of errors: (1) inherited errors that copy non-factual spans from the summary, and (2) errors introduced by imperfect QG models. While this latter set of errors may be eliminated by stronger QG models, inherited errors cannot be easily fixed.

vation, we hypothesize that such inherited errors in generated questions adversely affect the localization performance of automatic QA metrics by misclassifying factual spans.

First, we draw a clear distinction between (1) errors inherited from summaries, and (2) those introduced due to generation errors by the QG model. Figure 4 illustrates these two separate cases. We note that the latter set of errors can potentially be addressed by stronger QG models in the future. Our analysis in this section only studies the former set, i.e. inherited errors, as these will persist independent of the improvement in QA/QG models.

What percentage of questions are impacted by inherited errors? First, we determine the scope of the limitation introduced by inherited errors. Table 2 outlines how frequently generated questions contain inherited errors; we report these numbers only for the non-factual summaries as only these are affected by inherited errors. We define errors in a summary as inherited when a question copies at least one token that is annotated as non-factual. We use error type labels (extrinsic or intrinsic) present in the CLIFF and GD21 datasets to report separate numbers for these phenomena[4]. For the CLIFF dataset, we include world knowledge errors within the extrinsic type. In general, we observe that inherited errors are more common in QAFactEval compared to QuestEval; this can be attributed to the longer length questions generated by the former (see Appendix B for details).

[4]Generated questions can inherit both types of errors. In the tables in this section, "extrinsic error" denotes questions that inherit at least one extrinsic error, but "only intrinsic error" denotes questions that only inherit intrinsic errors.

QA Metric	Type of Inherited Error	GD21 XSum	CLIFF C/D	CLIFF XSum
QuestEval	extrinsic error	19.2	9.1	44.8
	only intrinsic error	25.7	17.3	3.6
QAFactEval	extrinsic error	39.1	11.1	93.1
	only intrinsic error	48.9	34.2	6.0

Table 2: Percentage of questions that inherit extrinsic and intrinsic errors from summaries. We only consider non-factual summaries, i.e., summaries containing at least one non-factual span in this table.

QA Metric	Type of Inherited Error	GD21 XSum	CLIFF C/D	CLIFF XSum
QuestEval	extrinsic error	3.1	93.9	30.5
	only intrinsic error	9.3	97.3	50.0
	no inherited error	15.5	98.7	56.0
QAFactEval	extrinsic error	7.7	65.4	63.8
	only intrinsic error	29.2	82.0	40.0
	no inherited error	29.3	92.8	65.4

Table 3: Percentage of factual spans correctly classified by QA metrics, i.e. $y_t = y_t^* = 1$. We use the same thresholds as for F1 scores in Figure 3. Results show that inherited errors lead to more erroneous classification as non-factual across all datasets.

Do inherited errors in generated questions hurt factuality prediction? To answer this, we zoom in on factual spans in generated summaries (we consider *both* factual and non-factual summaries here), and investigate how often these are erroneously classified as non-factual. We report this for three different scenarios: (1) w/ inherited extrinsic error, (2) w/ inherited intrinsic errors only, and (3) w/o any inherited error. Table 3 outlines our results. We observe that across all settings, factual spans with inherited errors in their corresponding questions are more likely to be erroneously classified as non-factual compared to those with no inherited errors. Between error types, we observe that extrinsic inherited errors tend to harm localization more than intrinsic errors.

Note that inherited errors are only observed for summaries that are *already non-factual*. Therefore, erroneous classification of factual spans as non-factual hurts span-level but does *not* hurt summary-level performance. In fact, Fabbri et al. (2022) show that longer questions (which typically inherit more extrinsic errors, but do not cause summary-level error) exhibit better summary-level performance compared to shorter questions (which can be under-specified and cause summary-level error). This indicates that there exists a trade-off in perfor-

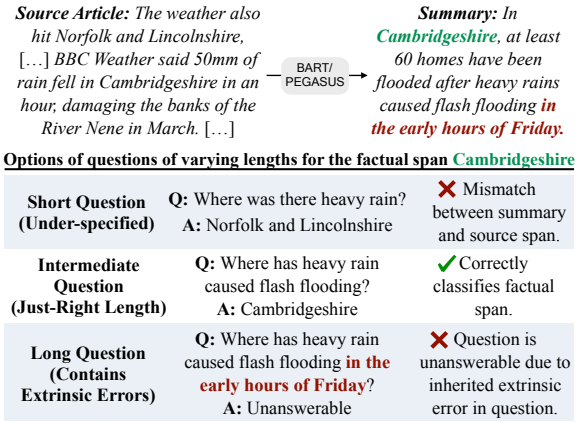


Figure 5: Specificity and lengths of generated questions affect the answers from QA models and confuse span-level evaluation. While "*Cambridgeshire*" is a factual span, questions with inappropriate specificity can cause QA models to make mistakes. It is impossible to know what length of question is "just right" during question generation.

mance between these different granularity levels.

Can we avoid inherited errors in generated summaries? Since we do not have prior knowledge of which spans in summaries contain factual errors, we cannot trivially ensure that generated questions do not inherit the errors. One possible strategy could be to generate very short questions that include minimal details from the summary to avoid inheriting non-factual spans from the summary. However, these may then suffer from being too under-specified. We illustrate this in Figure 5. Consider the factual span "*Cambridgeshire*". The shortest question in the figure "*Where was there heavy rain?*" is not under-specified for the summary, since it is the only place name in the summary. However, there are multiple possible answers in the source document, and QA models may *reasonably* answer "*Norfolk and Lincolnshire*", leading to erroneous classification of "*Cambridgeshire*" as non-factual. Therefore, **there exists a trade-off between under-specified (short) questions and over-specified (long) questions** and it is difficult to predict the optimal level of specificity. This problem cannot be addressed by improving QA models; an ideal QA model will return unanswerable to questions with inherited errors and will be still confused by under-specified questions. We explore this issue further using human question generation in Section 6.

6 Can Human QG Improve Localization?

In Section 5, we discussed how the number of inherited errors can be indirectly influenced by varying the length of generated questions and the accompanying trade-offs: longer questions are more likely to inherit errors but shorter questions may be under-specified. Here, we investigate this using perfect QG, i.e., replacing automatic QG with humans. We evaluate two aspects: (1) How does question length impact localization? (2) Does human QG improve localization?

6.1 Experiment Design

For each summary and candidate span pair (S, a_i), we obtain human-written questions of varying lengths and information content.[5] Then, we replace the QG module of QAFactEval with these human-written questions to study the effect of question length on error localization performance.[6]

Annotators generate 3 types of questions:

1. **Shortest possible question** such that given the question-summary pair, humans can unambiguously identify the correct span in the summary.
2. **Longest question** incorporating as much information from the generated summary as reasonably allowed, often including almost an entire summary sentence.
3. **Intermediate questions** with levels of information content between the above two extremes. We allow annotators to generate any number of such intermediate questions.

Annotation and Setup We conduct this experiment on 150 randomly selected summaries from the CLIFF dataset. For the CNN/DM subset, we only selected non-factual summaries, since CNN/DM contains a small number of non-factual spans. Human annotators manually generated 2,186 questions (please refer to Appendix D for details).[7] We use half of the summaries as validation sets.

We evaluate localization performance using 4 different length configurations for human-written questions: short, long, intermediate, and oracle.

[5] Question lengths could also be varied if we used distinct automatic QG models, but by choosing human QG, we avoid conflating the impact of varying question specificity/length with errors or other performance differences in models.

[6] We also considered using human QA; however, we found that the QA task is ill-defined for humans when questions themselves contain extrinsic errors. Fabbri et al. (2022) also show that QA performance has less impact on factuality.

[7] Human generated questions are provided at: https://github.com/ryokamoi/QA-metrics-human-annotation

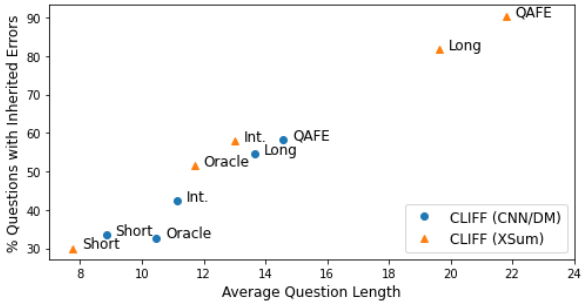


Figure 6: Statistics for questions generated by human annotators and QAFactEval (QAFE). "% Questions with Inherited Errors" is the percentage of questions that inherit non-factual spans from non-factual summaries. As expected, longer questions are more likely to inherit factual errors from the generated summaries.

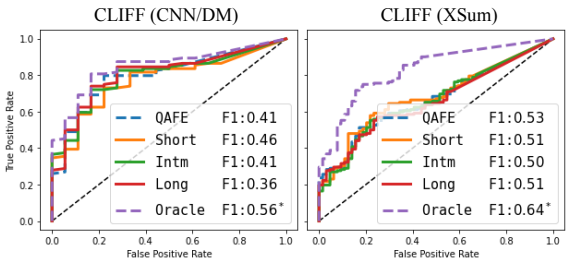


Figure 7: ROC curves and F1 scores (in legend) for span-level performances using human-written questions. These results show that no single question length configuration (except oracle) can outperform automatic QG. * denotes that improvement over QAFactEval is statistically significant (paired bootstrap test, p-value < 0.05).

For intermediate questions, evaluation is always done over three questions. We randomly subsample (or over-sample) from this set if more (or fewer) than three are available and report their average performance. If no intermediate question is written, we randomly sample from the other two categories. This only happens when the length difference between the shortest and longest questions is small. For the oracle setting, we report results using the question for each span that leads to the best localization performance. In other words, we use the highest scoring question for factual spans and the lowest for non-factual spans.

6.2 Results

Figure 6 outlines statistics for questions generated by human annotators and the QG model of QAFactEval (QAFE) generated for the same spans. As expected in Section 5, it shows that the percentage of questions that inherit non-factual spans in summaries increases with length. In this figure, we only analyze non-factual summaries since questions generated for factual summaries do not inherit errors. This result verifies our assumption and shows that we can analyze a trade-off between long questions that tend to inherit more non-factual spans from summaries and short questions with fewer inherited errors but can be under-specified.

Error Localization Figure 7 outlines the span-level localization performance for these different human question configurations and the QG model of QAFactEval. First, we notice that **human QG does not improve the localization performance of the QA frameworks**, with all three configurations exhibiting similar performance to the fully automatic QAFactEval (QAFE) model. **However, the *oracle* questions report significant improvement over QAFactEval**; this indicates that while there does exist an optimal length question for most spans, there isn't a clear pattern that can help select it during evaluation. We again note that it is not possible to select an optimal question length for each span without prior knowledge about errors in summaries. We conclude that the overall failure of human QG to improve over QAFactEval suggests that there exist fundamental issues with the QA-based factuality formulations which cannot be simply fixed by stronger QG models.

7 Discussion

Analysis in our paper suggests that QA-based metrics have fundamental problems which will be difficult to address in future work. **Our view is that future system designers should favor entailment-based approaches (Falke et al., 2019; Laban et al., 2022) as a result.** One reason for this is that successful QA-based approaches actually implement something similar to entailment. Both our analysis and Fabbri et al. (2022) show that we can improve summary-level performance by generating long questions to avoid underspecified questions. However, answering questions that contain almost all the information about a sentence can be regarded as a weak form of entailment evaluation: it assesses whether the question-answer pairs that include all information about the sentence are entailed by the original document. Compared to entailment, the answer comparison step can introduce difficulties and long questions may still lead to incorrect evaluation. Since this paper shows that QA-based metrics do not lead to interpretable, localizable judgments about errors, QA-based met-

rics do not seem to have any structural advantage over entailment-based metrics.

8 Related Work

Recent work (Fabbri et al., 2021) has shown that popular metrics like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) correlate poorly with human judgments of summary quality. Factuality, in particular, has been widely studied in recent summarization literature (Kryscinski et al., 2019; Falke et al., 2019), both from the perspective of identifying non-factual generations (Wang et al., 2020; Durmus et al., 2020; Goyal and Durrett, 2020) and improving the factuality of summarization models themselves (Kang and Hashimoto, 2020; Cao and Wang, 2021).

The majority of the work in factual evaluation has focused on summary-level metrics and is not capable of localizing errors within summaries. Recent work has decomposed factuality into summaries' dependency arcs (Goyal and Durrett, 2020) or semantic-graph representations (Ribeiro et al., 2022). These localization capabilities have several downstream applications like post-editing (Zhao et al., 2020; Chen et al., 2021), removing noisy training data (Nan et al., 2021; Goyal and Durrett, 2021), among others.

9 Conclusion

In this work, we show that although QA-based factuality metrics are motivated by error localization, in practice, they exhibit extremely poor localization capabilities. We provide a detailed analysis of the different issues in current metrics that hinder better localization performance. Finally, we run a human study to investigate whether human-level QG can fix some of these issues and conclude that there exist fundamental issues with the QA framework that cannot be simply fixed by stronger models.

10 Limitations

Given the lack of prior study in error localization of summarization evaluation, there is no large-scale dataset with token-level or span-level factuality labels. Constantly-evolving summarization models also mean that any such dataset would be come outdated in a fairly short time. However, we believe that the fundamental issues we discussed with QA metrics would persist across different summarization model outputs, despite our evaluation over a limited set.

Note that all our analysis is conducted on English language datasets and models of summarization, with a limited focus on newswire summaries. We believe that the issues identified here will transfer to other languages, but other domains such as dialogue or narrative summaries may exhibit substantially different types of factuality errors. These have not been studied as heavily in prior work, so likely new techniques and analysis will be needed for these settings.

Acknowledgments

We thank Juan Diego Rodriguez for helpful comments on this work. This work was partially supported by NSF grant IIS-2145280, a gift from Salesforce Research, and a gift from Amazon. The authors acknowledge the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for providing HPC resources used to conduct this research.

References

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

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

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the Original: Fact-Aware Neural Abstractive Summarization. *AAAI Conference on Artificial Intelligence (AAAI)*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Ping Chen, Fei Wu, Tong Wang, and Wei Ding. 2018. A Semantic QA-Based Approach for Text Summarization Evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving Faithfulness in Abstractive Summarization with Contrast Candidate Generation and Se-

lection. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5935–5941.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming Question Answering Datasets Into Natural Language Inference Datasets. *arXiv preprint arXiv:1809.02922*.

Daniel Deutsch and Dan Roth. 2021. Question-Based Salient Span Selection for More Controllable Text Summarization. *arXiv preprint arXiv:2111.07935*.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, 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.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

Tanya Goyal and Greg Durrett. 2020. Evaluating Factuality in Generation with Dependency-level Entailment. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 3592–3603.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Ranit Aharonov, and Sachindra Joshi. 2021. Using question answering rewards to improve abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 518–526, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, 28.

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

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

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.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Annual Meeting of the Association for Computational Linguistics*, page 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, page 74–81.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Mcdonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Annual Meeting of the Association for Computational Linguistics*.

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

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

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

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. *Journal of Machine Learning Research*, 21:1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. Factgraph: Evaluating factuality in summarization with semantic graph representations. *arXiv preprint arXiv:2204.06508*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Sentence Summarization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 379–389. Association for Computational Linguistics.

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

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

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

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.

Daniela Brook Weiss, Paul Roit, Ayal Klein, Ori Ernst, and Ido Dagan. 2021. QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 9879–9894.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *International Conference on Machine Learning (ICML)*, pages 11328–11339.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*, pages 1–43.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

A Additional Implementation Details

A.1 QuestEval

We use version 0.2.4 of the implementation by the authors with the recommended parameters (`task = 'summarization', do_weighter = False`).[8]

A.2 DAE

Our experiments use a trained model provided by authors (`DAE_xsum_human_best_ckpt`). This model is trained on a subset of XSum dataset (Maynez et al., 2020) that is distinct from examples in CLIFF and GD21.

A.3 Converting Tokens-level factuality scores to span-level scores

For the EM and the DAE baselines provide token-level factuality scores. We refer readers to the original DAE paper for details on how token-level scores are obtained (Goyal and Durrett, 2021). However, all our evaluation is designed to be at the span-level to align with QA metrics. To convert token-level scores to span-level, we annotate a span as non-factual if it contains any non-factual token.

A.4 Statistics for Ignored non-NP/NE tokens

The QA metrics do not evaluate the factuality of any token outside the boundary of a named entity of a noun phrase (discussed in Section 3.1). In Table 4, we show which kinds of tokens are ignored by the QA metrics but annotated as non-factual in our human annotated factuality datasets. Figure 5 provides an illustrative example of such ignored non-factual tokens in the different datasets).

	GD21	CLIFF	
	XSum	C/D	XSum
adposition	30.8	25.3	29.9
verb	24.4	23.7	23.5
auxiliary	15.2	10.0	17.1
punctuation	12.6	21.2	17.5
particle	6.7	3.9	4.5

Table 4: Statistics for non-factual POS-tags ouside the NP/NE boundaries and ignored by the QA metrics.

B Statistics of Generated Questions in QA-Based Metrics

Table 6 provides statistics for the generated questions from QuestEval and QAFactEval, highlighting the difference between these two metrics. On average, QAFactEval generates much longer questions. On the other hand, QuestEval generates a larger number of questions as it often generates multiple questions per candidate span.

[8] https://github.com/ThomasScialom/QuestEval

CLIFF (XSum)
An environmental permit **has been revoked following** a fire **at** a fuel recycling plant **in** Manchester**.**
CLIFF (CNN/DM)
A Japan Railway maglev train **hit** 603 kilometers **per** hour(374 miles **per** hour**) on** an experimental track **in** Yamanashi Tuesday**.** A spokesperson **said** the train **spent** 10.8 seconds **traveling above** 600 km **per** hour **, during** which it **covered** 1.8 kilometers **(** 1.1 miles **)** That **'s** nearly 20 football fields **in** the time it **took** you **to read** the last two sentences **.** Japan Railways **has been testing** their train **to figure out** the best operational speed **for** a planned route **between** Tokyo **and** Nagoya **scheduled to begin** service **in** 2027 **.**
GD21
high winds **and** heavy rain **have caused** flooding **at** a derbyshire theme park**, forcing** it **to close for** the weekend**.**

Table 5: Example of non-factual tokens outside NP/NE boundaries and ignored by the QA metric in factuality evaluation.

	Avg Question Length			Avg No. Questions		
	GD21	CLIFF		GD21	CLIFF	
	XSum	C/D	XSum	XSum	C/D	XSum
QAFactEval	26.3	16.2	21.1	4.9	10.0	3.9
QuestEval	16.4	11.6	13.9	8.3	20.5	7.5

Table 6: Statistics for the generated questions for the QuestEval and QAFactEval metrics.

C Performance Loss due to Span Filtering

In Section 3.1, we discussed that the current QA metrics do not evaluate non-NP/NE spans. These operational shortcomings prevent these metrics from providing a complete picture of error localization over all summary tokens. Here, we discuss another similar issue arising due to the question filtering step of the overall workflow (Step 3).

Although QA metrics select all NP/NE spans for evaluation during the candidate selection stage (Step1), some of these are filtered out if their corresponding question is of low quality: a_i for which $A(q_i, S) \neq a_i$ are also discarded from further evaluation. Since no errors are detected in these spans, they are considered to be factual.

We observed that this question filtering step removes around 30% of the NE/NPs in the QAFactEval framework. This implies that this metric only

Model	GD21 XSum	CLIFF CNN/DM	CLIFF XSum
EM	0.30	**0.27**	0.64^{D}
DAE	**0.32**	0.20^{E}	**0.78**
QE	0.19^{DE}	0.06^{DE}	0.45^{DE}
QAFE	0.21^{D}	0.13^{DE}	0.49^{DE}

Table 7: Span-level performance (F1 scores) over the subset of NP/NEs that are not discarded by either of the two QA metrics. ^{D} (or ^{E}) denotes that the performance difference with DAE (or EM) is statistically significant according to a paired bootstrap test (p-value < 0.05). Even under this generous setting, we observe that the QA metric show very poor performance.

evaluates 70% of the valid spans, potentially missing factual errors in the remaining NP/NEs. Note that these numbers are considerably lower for QuestEval (<5%) as it generates multiple questions for each candidate span and hence is more likely to include an acceptable question.

As this impacted the results in Figure 3, we can ask what is the performance of the QA metrics over spans that they **actually** evaluate for factuality? If this performance is high, we can reasonably assume that the QA metrics' localization capabilities can be improved through better question generation models. Table 7 outlines our results: we report F1 scores at the span-level when evaluating over the subset of NP/NEs that are evaluated by both the QE and QAFE models. Although the QA metrics report improved results over those reported in Figure 3, these are still low enough so as to not be useful for error localization in practical settings.

Figure 8 shows the corresponding ROC curves for these. These show similar trends: the performance of QAFactEval improves when evaluating on this subset, but is still not better than the baseline models. Interestingly, for CLIFF (CNN/DM), we found that most of this improvement comes from the subset of candidate spans whose questions contain only a small fraction of factual errors (Figure 9). This aligns with our analysis in Section 5 that showed that errors in generated questions inherited from non-factual summaries was one of the major reasons for performance degradation, since questions generated from summaries with small number of errors are expected to inherit fewer errors.

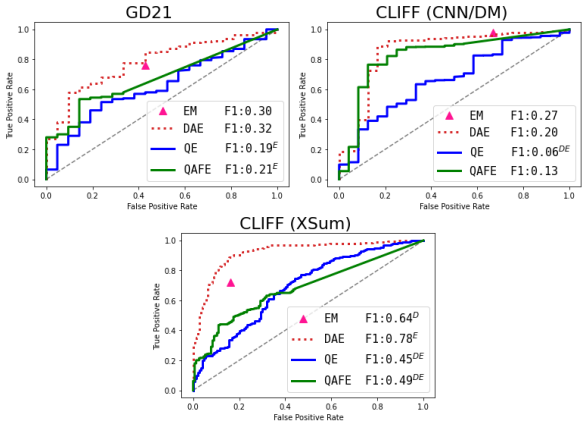
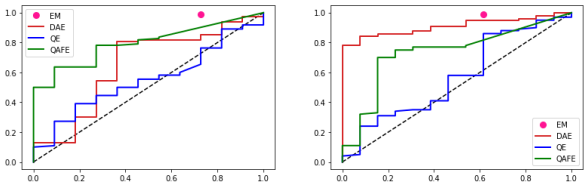


Figure 8: ROC curves for span-level performance on the subset of NE/NPs evaluated by all the QA metrics.



(a) Bottom 50% of the summaries according to error rate

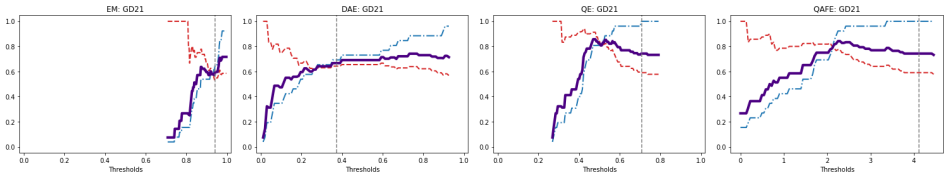
(b) Top 50% of the summaries according to error rate

Figure 9: Comparison of ROC curves for the span-level performance on CLIFF (CNN/DM). There shows results on the subset of NP/NEs actually judged by QA metrics. The graphs show that the performance of the QAFE is better on the subset of summaries with a smaller number of errors.

D Additional Details about Human QG Annotation

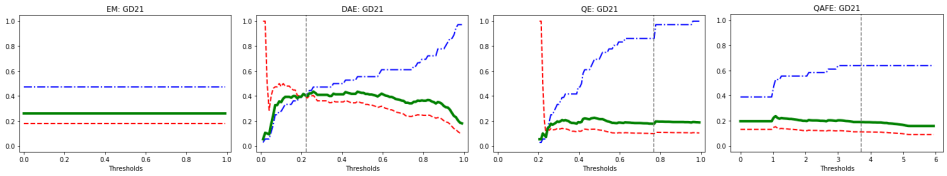
The human annotation in Section 6 was done by the authors of this paper. They were provided with summaries and extracted answer candidate spans. For spans that were judged to be invalid (e.g. "*it*"), they were asked to manually discard these spans. For all others, questions of varying lengths and specificity were written. See an example in Table 8. To aid in this question writing step, we also provide the corresponding QAFactEval questions. For the longest questions, we found that annotators often chose to build on these questions albeit with corrections to the structure and grammar of the automatic questions.

Table 9 shows the number of summaries, spans, and generated summaries annotated in our human QG experiments. We use half of the summaries as validation sets.



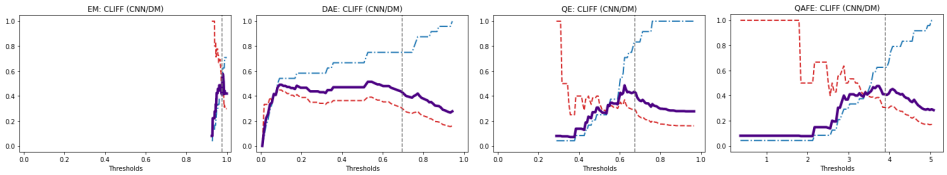
(a) EM Baseline (b) DAE (c) QuestEval (d) QAFactEval

Figure 10: Summary-level F1 performance on the GD21 test set at different thresholds



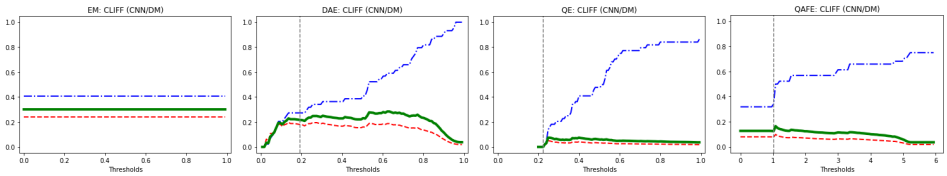
(a) EM Baseline (b) DAE (c) QuestEval (d) QAFactEval

Figure 11: Span-level F1 performance on the GD21 test set at different thresholds



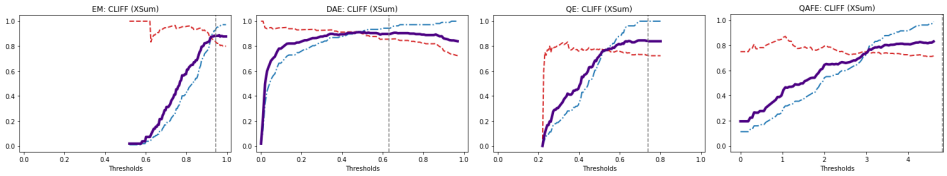
(a) EM Baseline (b) DAE (c) QuestEval (d) QAFactEval

Figure 12: Summary-level F1 performance on the CLIFF (CNNDM) test set at different thresholds.



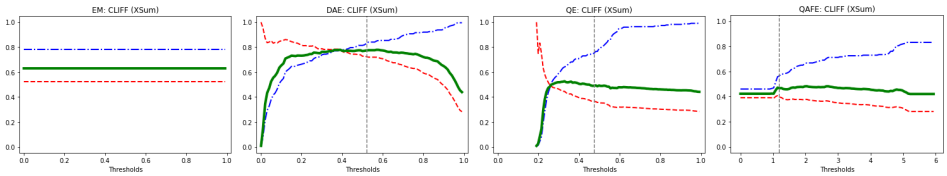
(a) EM Baseline (b) DAE (c) QuestEval (d) QAFactEval

Figure 13: Span-level F1 performance on the CLIFF (CNNDM) test set at different thresholds.



(a) EM Baseline (b) DAE (c) QuestEval (d) QAFactEval

Figure 14: Summary-level F1 performance on CLIFF (XSum) test set at different thresholds.



(a) EM Baseline (b) DAE (c) QuestEval (d) QAFactEval

Figure 15: Span-level F1 performance on the CLIFF (XSum) test set at different thresholds.

Summary	Plans to build a new hospital **in Somerset** have been given a **£3m** boost by the government.
Selected Span	a new hospital
Shortest	What does the plan propose to build?
Intermediate	What does the plan that has been given a boost propose to build?
Longest	What does the plan that has been given a **£3m** boost by the government propose to build?
QAFactEval	What plans to build **in Somerset** have been given a **£3m** boost by the government?

Table 8: Example of human-generated questions.

	CLIFF (CNN/DM)	CLIFF (XSum)
# Summary	30	120
# Span	323	470
# Questions	737	1449

Table 9: Number of summaries, spans, and generated questions annotated by human QG.

E Summary and Span Level Evaluation

We provide additional results for our experiments in Figure 3. Figure 10 to 15 show F1 scores, precision, and recall on test sets with different thresholds. These show that QA-based metrics cannot yield high precision at any threshold.