Explainable Assessment of Healthcare Articles with QA

Alodie Boissonnet[1], **Marzieh Saeidi**[2], **Vassilis Plachouras**[2], **Andreas Vlachos**[1,2]
[1]Department of Computer Science, University of Cambridge
[2]Facebook AI, London
{avmb2,av308}@cam.ac.uk, {marzieh,vplachouras,avlachos}@fb.com

Abstract

The healthcare domain suffers from the spread of poor quality articles on the Internet. While manual efforts exist to check the quality of online healthcare articles, they are not sufficient to assess all those in circulation. Such quality assessment can be automated as a text classification task, however, explanations for the labels are necessary for the users to trust the model predictions. While current explainable systems tackle explanation generation as summarization, we propose a new approach based on question answering (QA) that allows us to generate explanations for multiple criteria using a single model. We show that this QA-based approach is competitive with the current state-of-the-art, and complements summarization-based models for explainable quality assessment. We also introduce a human evaluation protocol more appropriate than automatic metrics for the evaluation of explanation generation models.

1 Introduction

The Internet has become an important source of medical advice. According to Rutten et al. (2019), in 2017, 74.4% of the US population first looked for health-related information on the internet, while only 13.3% of the population first asked a physician or healthcare provider. However, poor quality reporting, including misinformation, cherry-picking, exaggerations, etc., is often present online and can be a severe threat to public health. Recent events, such as the Covid-19 pandemic, demonstrate the necessity of developing quality assessment systems for healthcare reports to limit these harms. Fortunately, websites such as HealthNewsReview[1] critically analyze medical articles to identify poor quality reporting and improve the public discourse about healthcare. However, the manual review of medical news is a time-consuming task that would benefit from automated systems to scale up to the volumes needed in today's media ecosystem.

Story #1511
Criterion 1: Does the article adequately discuss the costs of the intervention?
Answer: Not Satisfactory
Explanation: There was no discussion of cost as there was in the competing AP story.

Criterion 2: Does the article adequately quantify the benefits of the treatment/test/product/procedure?
Answer: Satisfactory
Explanation: The story adequately quantified the benefits seen in the study that led to FDA approval.

Criterion 3: ...

Table 1: Example of an article evaluated by the HealthNewsReview website. Each article is evaluated according to ten criteria (two shown) and explanations are given to support the answers.

Assessing the quality of news articles has been the focus of numerous studies that tackle it as a text classification task (Louis and Nenkova, 2013; Chakraborty et al., 2016; Kryscinski et al., 2020). However, explanations for the predictions only recently started receiving attention, despite being necessary to convince the readers of such assessments. For instance, Dai et al. (2020) have built on the evaluation work conducted by the HealthNewsReview website (see Table 1) to automate article quality assessment in healthcare, but have only focused on articles classification, without providing explanations. Likewise, Wright and Augenstein (2021) have also studied exaggeration detection in healthcare as classification, but without explanations.

Beyond quality assessment, previous works have formulated textual explanation generation for classification as summarization (Atanasova et al., 2020; Kotonya and Toni, 2020). However such approaches suffer from a number of shortcomings when applied to the assessment of an article based on multiple criteria. As these approaches always

[1]https://www.healthnewsreview.org

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 1–9

output a single summary for a given input text, separate models must be trained to generate explanations for each classification label and evaluation criterion (e.g. reliability of sources, lack of information, etc.), as in the example given in Table 1. This considerably reduces the number of available training instances per model, because gold explanations of only one criterion at a time can be used for training, and it also requires developing and maintaining a model per criterion. Summarization-based models are also not appropriate to return an explanation for a label that is justified by the lack of information in the text (see criterion 1 in Table 1).

In this work, we develop an explainable quality assessment system for health news reports, and we evaluate it on the *FakeHealth* corpus (Dai et al., 2020). It differs from previous work as its explanation generation model is based on question answering (QA), which takes into consideration the definition of each evaluation criterion in the form of a question (see Table 1). This approach addresses the limitations of summarisation-based systems: it benefits from a larger training dataset consisting of instances from all criteria and labels at once, can better generate explanations regarding the absence of information, and requires training and maintaining a single model for all criteria.

We compare our approach for explanation generation against a summarization-based system inspired from Kotonya and Toni (2020). Our results show that both approaches are complementary and perform better in different cases. More specifically, summarization-based systems are more appropriate when relevant information is explicitly given in articles, while QA-based systems perform better when relevant information is missing.

Finally, evaluating generated explanations is not an easy task as we should consider both the structure and the sense of texts. Previous works used automatic metrics for the evaluation of explanations, which are known to be insufficient for abstractive text generation. Mani (2002) precisely insisted that assessing the readability and quality of a generated text requires human annotators as no automatic metric can achieve good performance on this task. Likewise, Kryscinski et al. (2019) have recently highlighted that automatic evaluation protocols, usually relying on ROUGE scores, correlate weakly with human judgement and fail to evaluate critical features, such as factual consistency. For this reason, we propose a new human evaluation protocol to assess the fluency, consistency, and factual correctness of the explanations, and we show that automatic metrics are not appropriate for this task.

2 Methodology

Our system starts with classifying articles according to ten evaluation criteria, then generates explanations using QA, taking into account the predicted classification labels. The purpose of the text classification step is to determine whether an article is satisfactory with respect to different evaluation criteria. We consider different options from the literature: logistic regression for its simplicity, BERT-based classification which is commonplace but truncates texts to 512 tokens, and a Longformer-based encoder model (Beltagy et al., 2020), which is able to deal with long input texts like those of our study. Both BERT and Longformer-based classifiers are pre-trained for a large classification task on a biomedical dataset, *PubMed*[2], then fine-tuned on the *FakeHealth* dataset. In line with Beltagy et al. (2020)'s recommendation, we use a classification objective for Longformer classifier, that places a global attention mask on a `[CLS]` token. This token aggregates the representation of the whole text at the beginning of the input text as shown in Table 6 in Appendix C.1, that gives an example of the encoding of input texts and shows the global attention mask of our model. Readers should refer to Beltagy et al. (2020) for further details about attention masks of Longformer models.

The second stage of the pipeline generates abstractive explanations for the previously predicted classes. As the QA approach takes into account the classes and the questions posed by criteria, we only need to train a single model, handling all criteria and classes. Following Soni and Roberts (2020), we have chosen to work with a Longformer-based encoder-decoder that we first train on the open-domain dataset *SQuAD v2.0* (Rajpurkar et al., 2018), and then fine-tune on *FakeHealth*. Because gold explanations in the *FakeHealth* dataset are abstractive, our model learns to write complete explanations despite the pre-training step on *SQuAD* whose explanations are spans of phrases. Even though we always use the same ten questions (shown in Table 9 in Appendix C.2) for fine-tuning

[2] https://deepai.org/dataset/pubmed

and evaluation, this approach differs from query-focused summarization because of its ability to generate explanations for information missing from the article which a summarization system cannot handle. We use the QA objective introduced by Beltagy et al. (2020) for Longformer that places a global attention mask on all question tokens (see Table 6 in Appendix C.1), and we feed our model with the article, the criterion, and the class prediction. During training, we use the gold classes of articles to generate explanations, as generating post-hoc explanations for incorrectly predicted labels would not be meaningful.

Following recent previous work on explainable fact-checking in healthcare by Kotonya and Toni (2020), we implement a baseline for the explanation generation task, based on summarization. Because such a system does not take into account the criteria definitions in its input, it cannot combine all criteria together as it would always produce the same explanation for all criteria. Therefore, this approach requires training independent models for each class within a criterion, which results in 30 models (10 criteria $\times$ 3 classes) in the case of the *FakeHealth* dataset. We use here a summarization objective for the Longformer model, that applies a global attention mask to the very first token of input texts (see Table 6 in Appendix C.1 and Beltagy et al. (2020)).

3 Human evaluation of explanations

Unlike previous works that assess generated text with automatic metrics, we design a human evaluation protocol that assesses four aspects of explanations: their fluency, consistency, factual correctness, and whether they are indicative of the label that they are supposed to explain. An explanation is considered fluent if it sounds natural, and consistent if it does not contradict itself, include repetitions, or information that is not mentioned in the article. The factual correctness criterion looks for incorrect facts, contradictions with respect to the article, or hallucinations. Finally, generated explanations should allow a human judge to infer correctly the label they are meant to explain.

We conducted two pilot studies in order to assess the quality of our guidelines. As reported in Table 2, Pilot 1 brought to light the ambiguity of the initial version of the guidelines, while Pilot 2 reached higher inter-annotator agreement scores. The new version of the guidelines is more detailed than the first one and provides some examples of what is expected. For instance, instead of asking if an explanation is fluent, the new guidelines specify that explanations should be rated as fluent if they sound natural and their syntactic structure is correct. Thus, the sentence "it's sunny but it's sunny" should not be considered as fluent, while "it's sunny but it's not sunny" should be considered fluent despite the contradiction, which is judged negatively under consistency.

The final guidelines used for the evaluation in Section 5 are fully detailed in Appendix B. In Table 2, the consistency criterion is missing as it was added after Pilot 2.

	Fluency	**Factual correctness**	**Guessed class**
Pilot 1	-0.12	0.29	0.76
Pilot 2	0.46	0.49	0.58

Table 2: Inter-annotator agreement scores (averages of Cohen Kappa scores) of the two pilot studies.

Criterion	**Not S.**	**S.**	**Not A.**
1	1431	495	370
2	1505	768	**23**
3	1413	717	**166**
4	1445	848	**3**
5	**286**	1921	**89**
6	1135	1147	**14**
7	1120	1063	**113**
8	538	1457	301
9	672	1543	**81**
10	391	1771	**134**

Table 3: Distribution of articles in each class per criterion. These numbers combine both the *HealthRelease* and *HealthStory* datasets.

4 Data

We evaluate our QA and summarization-based models on the *FakeHealth* corpus of health news articles, released by Dai et al. (2020). Each article in the dataset was evaluated by at least two experts, according to ten criteria that assess diverse aspects such as "the overclaiming, missing of information, reliability of sources and conflict of interests" (Dai et al., 2020). Dai et al. (2020) found zero to a minor positive correlation between the criteria, which justifies the relevance of all of them. These criteria are reported in Table 9 in Appendix C.2.

For each criterion, articles are annotated with one of three labels, *Not Satisfactory*, *Satisfactory*, and *Not Applicable*, and a textual explanation that justifies the assigned label, as shown in Table 1. The label distribution across criteria is highly unbalanced, *Not Applicable* instances being the rarest. For example, criteria 2, 4, and 6 have at least 65 times more *Not Satisfactory* instances than *Not Applicable* ones (see Table 3).

5 Results

5.1 Quality assessment per criterion

We compare Longformer-based, BERT-based, and Logistic Regression models for the quality of the classification task via their macro F_1-scores for each criterion. Table 4 shows that our Longformer-based models perform the best due to their ability to encode longer texts. The Logistic Regression models also achieves great performance despite its simplicity, but this must be qualified as classes are highly unbalanced and Logistic Regression mostly predicts the dominant class. An analysis broken down by criterion also highlights that all models perform unevenly across criteria. This suggests that some criteria are harder to handle, notably, those requiring external knowledge or subjective judgment (e.g. criterion 5 asking whether articles commit disease-mongering).

We also tried to build a single Longformer-based model handling all classes at once using a QA-based approach that treats criteria as questions and predicts labels, but it performed poorly. We suspect that we have poor results because we perform a classification task with a QA-based model.

5.2 Explanation generation

Table 5 reports the overall performance of both summarization and QA-based approaches for the explanation generation task only. These results show that the QA-based approach performs better than, or as well as, the baseline system. Both approaches achieve similar performance in terms of consistency and factual correctness, but the QA approach produces explanations that are more fluent and that indicate the correct label more often. Table 7 in Appendix C.2 provides some examples of the generated explanations. In these tables, gold explanations correspond to the explanations written by health expert in the *FakeHealth* dataset.

An analysis per class (see Table 5) reveals that the summarization approach performs better for the *Satisfactory* class, while the QA approach performs better for the *Not Satisfactory* and *Not Applicable* classes. This can be explained by the fact that *Satisfactory* articles include the relevant information to the criteria and require models to reuse this information to generate explanations, thus resembling summarization. On the other hand, for the *Not Satisfactory* class, models need to point out missing information and this is naturally harder for a summarization model, but easier for a QA-based one that can generate text about missing information. Finally, the *Not Applicable* class suffers mainly from having very few instances for training (see Table 3). With a single model, the QA approach is able to overcome this issue and generate better explanations.

	Longformer	**BERT**	**LogReg**	**From gen. expl.**
Criterion 1	**0.67**	0.63	0.59	0.61
Criterion 2	**0.43**	0.42	0.40	0.30
Criterion 3	0.52	**0.55**	0.46	0.45
Criterion 4	0.40	0.42	0.36	**0.61**
Criterion 5	0.35	0.30	**0.37**	0.33
Criterion 6	0.42	0.39	0.37	**0.60**
Criterion 7	0.35	0.37	0.36	**0.40**
Criterion 8	0.57	**0.59**	0.49	0.46
Criterion 9	**0.40**	0.37	0.37	0.34
Criterion 10	**0.45**	**0.45**	0.36	0.24
Mean	**0.46**	0.44	0.41	0.43

Table 4: Macro F_1-scores of our different classifiers for each criterion. The last row *Mean* gives the average performance of each model across criteria. The column *From gen. expl.* corresponds to the classification task conducted from generated explanations, as described in Section 5.3.

To achieve the best performance, the previous results suggest combining both systems and using the summarization-based system for *Satisfactory* instances, and the QA-based system for all others. With this combination, 81% of explanations are fluent, 76% consistent, 57% factually correct, and 85% indicate correct labels. The pretty low factual correctness of explanations can be explained by the severeness of guidelines that ask annotators to rate an explanation as factually incorrect as soon as at least one detail is incorrect, regardless of the correctness of all other details.

5.3 Predicting classes from generated explanations

To further test our methodology, we run an experiment in which we first generate explanations,

	Fluency		Consistency		Factual correctness		Correct class		Count
	Sum.	QA	Sum.	QA	Sum.	QA	Sum.	QA	
All classes	74.5	**80**	72.5	72.5	52.5	52	85	**86**	-
Not S.	73.2	**83.5**	67	**73.2**	42.3	**48.5**	87.6	**89.7**	97
S.	**79.6**	76.3	**80.6**	73.1	**63.4**	53.8	**86**	82.8	93
Not A.	40	**80**	50	**60**	50	**70**	50	**80**	10

Table 5: Results of the evaluation of the summarization and QA-based systems per class (as percentages).

and then classify articles from the predicted explanations. We use the same approach as before, i.e. a Longformer-based model with a QA objective fine-tuned on *FakeHealth* articles for explanation generation, and a Longformer-based classifier fine-tuned on predicted explanations. Results are reported in Table 4 and show that classifying articles before generating explanations, achieves better performance. This finding is not surprising as the explanation generation model is influenced by dominant classes and ignores minority classes. Wrong explanations propagate then to the classification task and are responsible for incorrect labels. However, the classification model built from generated explanations performs very well for criteria 4 and 6. Yet, these results should be considered with caution, as classes for these criteria are highly unbalanced (with respectively 3 and 14 instances in the *Not Applicable* class) and the model predicts most of the time the majority class. This ablation study corroborates the recommendations of Kotonya and Toni (2020) and Mani (2002).

5.4 Automatic v. human evaluation

Finally, we investigate the correlation between human judgement and automatic metrics used in previous works (Ermakova et al., 2019), including ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) scores. Table 8 in Appendix B.3 reports the correlation coefficients between all metrics. Using Kendall's Tau, we find that all these correlations are very low, at most 0.11 with ROUGE scores and 0.07 with the BLEU score. This finding was expected as most of the automatic metrics focus on word overlap, which makes it difficult to check the grammatical and syntactic correctness of explanations, as well as their factual consistency. This conclusion echoes Kryscinski et al. (2019)'s work on automatic evaluation protocols.

6 Conclusion and discussion

In this work, we propose a new QA-based approach to generate explanations for quality assessment systems. This approach allows us to build a single model, able to generate explanations for different criteria and classes, by taking into account the questions related to criteria. We have shown that the QA-based system is competitive with the summarization-based one, and that they are complementary. Notably, the QA-based approach is more appropriate when the relevant information is not explicitly given in articles or for small classes. As for the classification task, Longformer-based models perform best thanks to their ability to deal with long input texts. Finally, we have highlighted that automatic metrics, such as ROUGE, correlate very weakly with human judgment when it comes to evaluating explanation generation models. This paper could serve as a starting point to explore the use of QA models for explainable article assessment.

Acknowledgements

Andreas Vlachos's work at the University of Cambridge is supported by by the ERC grant AVeriTeC (GA 865958) and the EU H2020 grant MONITIO (GA 965576).

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations.

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

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16.

Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. *CoRR*, abs/2002.00837.

Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. 2019. A survey on evaluation of summarization methods. *Information Processing Management*, 56(5):1794–1814.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims.

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.

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.

Annie Louis and Ani Nenkova. 2013. What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Inderjeet Mani. 2002. Summarization evaluation: An overview.

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.

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.

Lila J. Finney Rutten, Kelly D. Blake, Alexandra J. Greenberg-Worisek, Summer V. Allen, Richard P. Moser, and Bradford W. Hesse. 2019. Online health information seeking among us adults: Measuring progress toward a healthy people 2020 objective. *Public Health Reports*, 134(6):617–625. PMID: 31513756.

Sarvesh Soni and Kirk Roberts. 2020. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5532–5538, Marseille, France. European Language Resources Association.

Dustin Wright and Isabelle Augenstein. 2021. Semi-supervised exaggeration detection of health science press releases.

A Ethical concerns

The ethical concerns of this work are two-fold. First, readers must be aware that such a deep learning model is prone to make mistakes, as evidenced by the results of the experiments we did (see Section 5). Outputs should be treated as an indication or recommendation, rather than the ground truth.

Secondly, our QA-based approach needs to train a single model, by comparison with the summarization-based one that requires 30 models. Having a single model reduces the pressure on computing resources and consequently, on the environment. It also makes the model easier to maintain.

B Human evaluation

B.1 Definition of the evaluation guidelines

To design our human evaluation protocol, we conducted two pilot studies with the same two annotators. To begin with, the first study gathered three annotators who evaluated all explanations generated for the same six articles (three releases and three stories, which results in 60 explanations in total) with the baseline system for explanation generation. They were asked to determine if explanations were written in fluent English, consistent, factually correct, and which classes were suggested by explanations. This evaluation task combined both intrinsic and extrinsic methods to have a complete overview of models' performance, and we assessed to what extent annotators agreed on the evaluation task by looking at inter-annotator agreement scores computed with the Cohen Kappa score. It resulted in a high disagreement among annotators (see Table 2): annotators 1 and 2 even seemed to disagree on the fluency criterion. An in-depth exploration of their annotations revealed that they never agreed when one of them judged that an explanation was not fluent. These low inter-annotator agreement scores seem therefore to be caused by unclear guidelines.

For this reason, more detailed guidelines about the fluency and factual correctness of explanations

were defined, and another pilot study was intended to validate them. It gathered two of the three previous annotators, who evaluated all explanations generated for the same five articles (two releases and three stories) with whether the baseline or the QA-based system. We reduced the number of articles to evaluate as evaluation tasks are time-consuming and five articles, resulting in 50 explanations, are enough to validate guidelines. This second evaluation task achieved a much higher inter-annotator agreement reported in Table 2 and confirmed the new evaluation guidelines. However, the agreement score for the guessed classes slightly decreased between the first and second evaluation task. An analysis of annotations highlighted that some criteria could be ambiguous. For example, criterion 5 wonders if articles commit disease-mongering, and if they do, they should be rated as *Not Satisfactory* because it implies that they are less reliable. Consequently, a detailed description of each criterion, extracted from HealthNewsReview's website, has been given to annotators for the last evaluation task to raise all ambiguities.

B.2 Final guidelines

Based on the outcome of the pilot studies, annotators were given the following guidelines:

- Fluency: Is the generated explanation written in fluent English? An explanation should be considered non-fluent if it does not sound natural or its structure is not correct (e.g. paragraphs title). Words case (uppercase or lowercase) should not be taken into account. For example, "it's sunny but it's sunny" should be considered as non-fluent, but "it's sunny but it's not sunny" should be considered as fluent. Likewise, "intro: it's sunny, results: it's sunny, conclusion: it's sunny" should be considered as non-fluent (inappropriate structure).

- Consistency: Is the generated explanation consistent? An explanation should be considered inconsistent if it includes contradiction, repetition, extra information. For example, "it's sunny but it's sunny" should be considered as consistent, but "it's sunny but it's not sunny" should be considered as non-consistent.

- Factual correctness: Are the details (numbers, names, facts, etc.) included in the generated explanation correct? Explanations that contain incorrect facts, contradictions, or hallucinations should be evaluated as not satisfactory; but whether or not the factual details are related to the question should not be taken into consideration.

- Suggested class: According to the generated explanation, how would you classify the article? (*Not Satisfactory*, *Satisfactory*, *Not Applicable*, *Can't tell*) A *Can't tell* class has been added if generated explanations do not help classify articles. A description of what was expected for each criterion was given to annotators to raise all ambiguities. It was taken from the HealthNewsReview website from which explanations had been extracted. The inferred classes are considered correct if it matches the gold classes of articles.

The consistency criterion has been added after the two pilot studies, so we have not evaluated the inter-annotator agreement for it. However, the corresponding guidelines have been defined and detailed similarly to the other evaluation criteria to raise any ambiguity for annotators.

For the real evaluation task, annotators have evaluated ten different articles each. They were the same annotators as for pilot studies, so their inter-annotator agreement was high and we were able to evaluate more articles with great confidence in annotations.

B.3 Correlation with automatic metrics

Table 8 reports the correlation scores between human judgement and automatic metrics used in previous works (Ermakova et al., 2019), including ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) scores. Using Kendall's Tau, we find that all these correlations are very low, at most 0.11 with ROUGE scores and 0.07 with the BLEU score.

C Model

C.1 Model's Attention

For the Longformer model, Beltagy et al. (2020) defines different global attention masks according to the task to complete. For classification, the `[CLS]` token of input texts receives global attention. For a QA task, the global attention mask is applied to all question tokens, while it is applied to the very first token of input texts for a summarization task. Table 6 illustrates these different attention masks.

C.2 Example of models' outputs

Question-Answering objective

<s> Does the story adequately discuss the costs of the intervention? </s> Satisfactory </s> Triumph for Drug to Straighten Clenched Fingers </s> About one in 20 men is estimated to have Peyronie's, but figures are not precise because people with the condition tend not to discuss it publicly. ... </s>

Summarization objective

<s> Triumph for Drug to Straighten Clenched Fingers </s> About one in 20 men is estimated to have Peyronie's, but figures are not precise because people with the condition tend not to discuss it publicly. ... </s>

Classification objective

[CLS] Triumph for Drug to Straighten Clenched Fingers </s> About one in 20 men is estimated to have Peyronie's, but figures are not precise because people with the condition tend not to discuss it publicly. ... </s>

Table 6: Example of the inputs and global attention masks of the Longformer model for the different training objectives for HealthStory 1613, Criterion 1. The global attention is put on the tokens highlighted in grey.

HealthRelease #528, Criterion 5, Satisfactory

Gold explanation

There is no disease mongering in this release.

Generated explanation – QA-based system

There is no disease mongering in this news release. Obesity is a serious health issue, and the release provides some context about the prevalence of obesity in the US.
Rated as fluent, consistent, factually incorrect, and suggesting a *Satisfactory* label

Generated explanation – baseline system

There is no disease mongering in the news release. The release provides some context about the prevalence and severity of obesity and its consequences.
Rated as fluent, consistent, factually incorrect, and suggesting a *Satisfactory* label

---

HealthStory #1619, Criterion 1, Satisfactory

Gold explanation

The story notes that a 5.29-ounce tube of the product retails for $39.99 at drugstores.

Generated explanation – QA-based system

The story notes that the cream is available at Sephora.com for $69.99. The story could have done a better job of comparing the cost of the cream to other products on the market.
Rated as fluent, consistent, factually incorrect, and suggesting a *Satisfactory* label

Generated explanation – baseline system

The story states that the cream "takes the redness out of new stretch marks." That's good enough for the story, but it would have been better if the story would have mentioned the cost.
Rated as fluent, consistent, factually correct, and suggesting a *Not Satisfactory* label

Table 7: Examples of explanations generated with our QA-based and summarization-based baseline systems.

	Fluency	Consistency	Factual correctness	Correct class	Can't tell class
ROUGE-1 Precision	0.09	0.04	0.03	0.08	-0.07
ROUGE-1 Recall	-0.02	-0.08	-0.05	-0.04	-0.00
ROUGE-1 F1	0.01	-0.05	-0.01	0.00	-0.04
ROUGE-2 Precision	0.08	0.05	0.04	0.09	**-0.11**
ROUGE-2 Recall	0.04	-0.02	-0.01	0.04	-0.09
ROUGE-2 F1	0.06	0.01	0.01	0.07	**-0.11**
ROUGE-L Precision	0.10	0.08	0.05	0.09	-0.09
ROUGE-L Recall	0.01	-0.04	-0.03	-0.01	-0.03
ROUGE-L F1	0.06	0.03	0.02	0.06	-0.08
BLEU	-0.01	-0.07	-0.04	-0.01	-0.03
Length ratio	0.09	0.08	0.05	0.08	-0.06
Cosine similarity	0.08	-0.01	0.03	0.06	-0.05
Euclidean distance	-0.04	0.01	-0.04	-0.02	0.03

Table 8: Correlation between human and automatic evaluation metrics (Kendall Tau correlation coefficient).

Criterion	Question
Criterion 1	Does it adequately discuss the costs of the intervention?
Criterion 2	Does it adequately quantify the benefits of the treatment/test/product/procedure?
Criterion 3	Does it adequately explain/quantify the harms of the intervention?
Criterion 4	Does it seem to grasp the quality of the evidence?
Criterion 5	Does it commit disease-mongering?
Criterion 6	Does the story use independent sources and identify conflicts of interest? / Does the news release identify funding sources & disclose conflicts of interest?
Criterion 7	Does it compare the new approach with existing alternatives?
Criterion 8	Does it establish the availability of the treatment/test/product/procedure?
Criterion 9	Does it establish the true novelty of the approach?
Criterion 10	Does the story appear to rely solely or largely on a news release? / Does the news release include unjustifiable, sensational language, including in the quotes of researchers?

Table 9: Datasets' criteria.