HaRiM$^+$: Evaluating Summary Quality with Hallucination Risk

Seonil Son, Junsoo Park, Jeong-in Hwang, Junghwa Lee, Hyungjong Noh, Yeonsoo Lee
NCSOFT NLP Center
{deftson,junsoopark,jihwang,jleehhh0217,nohhj0209,yeonsoo}@ncsoft.com

Abstract

One of the challenges of developing a summarization model arises from the difficulty in measuring the factual inconsistency of the generated text. In this study, we reinterpret the decoder overconfidence-regularizing objective suggested in (Miao et al., 2021) as a hallucination risk measurement to better estimate the quality of generated summaries. We propose a reference-free metric, HaRiM$^+$, which only requires an off-the-shelf summarization model to compute the hallucination risk based on token likelihoods. Deploying it requires no additional training of models or ad-hoc modules, which usually need alignment to human judgments. For summary-quality estimation, HaRiM$^+$ records state-of-the-art correlation to human judgment on three summary-quality annotation sets: FRANK, QAGS, and SummEval. We hope that our work, which merits the use of summarization models, facilitates the progress of both automated evaluation and generation of summary.

1 Introduction

Although recent state-of-the-art summarization models have achieved remarkable performances (Lewis et al., 2020; Raffel et al., 2020; Zhang et al., 2020), appropriate metrics for measuring faithfulness of the generated summaries are still needed. The practice of measuring performance in the summarization task heavily relies on the N-gram matching based metric, ROUGE (Lin, 2004). Reportedly, ROUGE barely satisfies more than indicating lexical similarity (Maynez et al., 2020) and does not consider semantic dimensions of the generation, which current research needs of.

There have been numerous attempts to come up with faithfulness evaluation metrics (Novikova et al., 2017; Peyrard, 2019). Neural-based metrics have demonstrated good performances in estimating the factual consistency of a summary-article pair with semantic entailment (Kryscinski et al., 2020; Goyal and Durrett, 2020), question-answering framework (Wang et al., 2020; Scialom et al., 2021, 2019), and text generation (Yuan et al., 2021; Xie et al., 2021). Most of the model-as-a-metric approach generally requires fine-tuning or complicated pipelines. Consequently, evaluating generated texts with recent model-as-a-metric methods has become cumbersome.

With the increased demand for faithful generation models, it has come to a lot of attention on reformulating training objectives for purported for this (Zhang et al., 2022; Liu et al., 2022a; Holtzman et al., 2018). We focus on the training objective suggested in (Miao et al., 2021), which directly targets hallucination problems in generating sentences given a source context. Miao et al. suggest that an overconfident decoder causes hallucination since the model excessively pays attention to the previously generated tokens over the source context which is in line with (Bowman et al., 2016).

In this paper, we reinterpret the decoder overconfidence regularization term from (Miao et al., 2021) as *hallucination risk* and recompose the objective to be practical for summary quality evaluation in various aspects. Unlike other recent metrics (Yuan et al., 2021; Xie et al., 2021), our metric, HaRiM$^+$, detects hallucination in summary texts and evaluate their quality with the help of log-likelihood of summarization models. Also, HaRiM$^+$ does not require complicated pipelines, further training, or modification of the generation model in use.

We conduct experiments to verify the effectiveness of our metric on several summary quality estimation benchmarks. We test HaRiM$^+$ on FRANK, annotation sets from QAGS, and SummEval, which provides multiple aspects of summary-quality judgements accompanied by summarization system outputs. Through quantitative and qualitative experiments, we demonstrate the robust performance of our metric HaRiM$^+$, present the analysis of its inductive bias, and potential ex-

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 895–924
November 20–23, 2022.

tension.

2 Related Works

2.1 Evaluation of Text Generation

Automatic evaluation of generated text, despite its importance, has long relied on token-wise comparison against a reference target, and has been insufficient for reliably reflecting correctness and consistency. Most commonly used metrics, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), are N-gram based metrics that compare token overlaps between candidate and reference texts. Model based metrics such as BERTScore (Zhang et al., 2019) use BERT representation of tokens, but such approaches have exhibited low correlation with human judgments of correctness for summarization datasets (Wang et al., 2020).

As text generation models improve, sequence-to-sequence text generation models are increasingly being used for text quality evaluation. BARTScore (Yuan et al., 2021) leverages the generation model's ability to assign higher probability to reference source-target pairs. PRISM (Thompson and Post, 2020) is a multilingual translation model that is used as a reference-to-candidate paraphraser. COCO (Xie et al., 2021) measures quality by estimating the effect of the language prior in text generation that contributes to hallucination. The idea of using text generation models to estimate the log-likelihood of candidate sequences is conceptually simple yet has shown to be effective in evaluating text quality. Our approach follows this line of research, but aims to improve the judgments of the consistency of the generated summary by adding a hallucination risk term.

2.2 Hallucination Detection in Summarization

Numerous works have addressed the need for an automatic way of detecting hallucination in generated summaries. This can be accomplished by reformulating detection problem into auxiliary tasks. Textual entailment-based approaches consider the summary hallucination problem as a natural language inference (NLI) task, and leverage NLI classification models to score candidate summaries (Falke et al., 2019). QA-based approaches employ question generation and question answering models to generate questions from the candidate summary and to check the answerability of the question, respectively (Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2019, 2021). (Goyal and Durrett, 2020) propose to utilize dependency parser to classify whether each dependency arc is hallucinated. QA-based approaches resemble the PYRAMID method (Nenkova and Passonneau, 2004) and its automated descendants (Harnly et al., 2005; Passonneau et al., 2013; Gao et al., 2019) from a content selection perspective.

More direct approaches attempt to use models that are trained to distinguish artificially generated set of negative summaries. Kryscinski et al. augments factual article-summary pairs to generate data for training a classification model. Zhou et al. employs a token-level prediction model to be trained on generated hallucination data. All of the above methods require the generation of additional datasets and the training of auxiliary models. In contrast, our approach only requires an off-the-shelf abstractive summarization model that needs no further training, and eliminates the need for preparing additional data.

3 Method

We describe the logic behind *margin-based token-level objective* (Miao et al., 2021), and reinterpret it as *hallucination risk*. We then propose modifications to re-formulate the original objective to be feasible for evaluating text quality.

3.1 Hallucination Risk Measurement (HaRiM)

In encoder-decoder architectures, having the decoder relying too much on the decoder's context and less on the encoder's is a long known problem (Bowman et al., 2016). Miao et al. introduced *margin-based token-level objective* as a regularization term that prevents the decoder from focusing too much on the decoder-side context. Considering that hallucination refers to erroneous generation irrelevant to the source context, the regularization term can be reinterpreted as *hallucination risk*. For source input text X and target text $Y = \{y_0, y_1, ..., y_L\}$, the term HaRiM is defined as:

$$\text{HaRiM} = \frac{1}{L}\sum_{i=0}^{L}(1-p_{s2s})(1-(p_{s2s}-p_{lm})) \quad (1)$$

where p_{s2s} and p_{lm} represent the token-likelihood of the sequence-to-sequence model (S2S) and that of the auxiliary language model (LM) respectively,

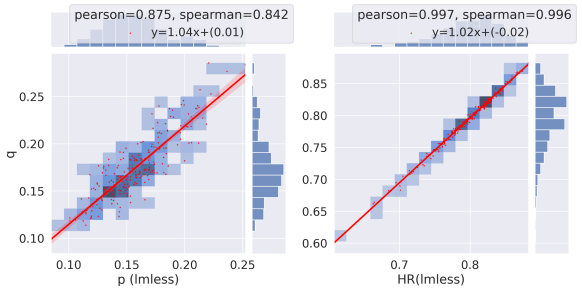

Figure 1: Effects of replacing the auxiliary language model ($q(y_i|y_{<i})$) with an empty-sourced encoder-decoder model ($p(y_i|y_{<i};\{\})$. **Left** compares the values of p_{lm}, and **Right** compares the HaRiM values. The values are calculated on the summary-article pairs in FRANK benchmark. The high correlation of HaRiM suggests that the effect of replacement is minimal.

and are defined as:

$$p_{s2s} = p(y_i|y_{<i}; X),\ p_{lm} = q(y_i|y_{<i}) \quad (2)$$

The S2S measures the probability of a target sequence with the knowledge of the encoder input X, while the LM does the same without X. The value of HaRiM increases as the p_{lm} overwhelms p_{s2s}. The value is weighted inversely by the S2S likelihood, thus maximizing when the S2S likelihood minimizes.

As described in the original paper, Equation 1 is one of many ways of implementing the *hallucination risk* using token likelihoods. However, after exploring many variations[1], we decide that the form in Equation 1 works best for our purpose of quality estimation.

3.2 Recomposing HaRiM for Feasible Evaluation

Replacing Auxiliary Language Model with Empty-Sourced Encoder-Decoder

One of the challenges in applying hallucination risk to text evaluation is the requirement of the auxiliary language model ($q(\cdot)$ in Equation 2) for the risk computation. Miao et al. formulate the language model as an auxiliary decoder-only model that is jointly trained with the main encoder-decoder of the S2S model. However, when using an off-the-shelf summarization model for summary quality evaluation, this approach is infeasible because it needs a language model that should have been trained jointly with the summarization model, especially on a limited summarization dataset that can be insufficient for training a language model. To avoid the joint training of language model, one can consider using a pre-trained language model to replace the auxiliary model. However this approach is also infeasible because the tokenization and vocabulary of the language model must match the ones of the S2S model.

Instead we consider re-purposing the entire encoder-decoder from the summarization model itself as a language model. In this way, the LM model is simply the S2S model itself, but works as an LM when it receives an empty source text (denoted as $\{\}$) as the encoder input. This eliminates the need for an additional model, and automatically solves the tokenization and vocabulary issue as well. Thus we replace the p_{lm} from auxiliary language model likelihood ($q(\cdot)$) to empty-sourced S2S likelihood as the following[2]:

$$p_{lm} = p_{s2s}(y_i|y_{<i};\{\}) \quad (3)$$

We test the validity of such modified use of S2S model as the LM model when calculating the hallucination risk. We compare the hallucination risk value when replacing p_{lm} from auxiliary language model to empty-sourced S2S. The results in Figure 1 show that hallucination risk HaRiM calculated with empty-sourced S2S is almost perfectly linear with the counterpart computed with the auxiliary model ($\rho = .997$), thus p_{lm} is replaceable as the Equation 3 in computation of HaRiM.[3]

Accompanying HaRiM with Log-likelihood (HaRiM$^+$)

A broad range of factors for text quality estimation makes evaluation task hard because it varies according to the generation task. An implicit way of measuring overall generation quality is to use token likelihood of high-performing text-generation models as reported in (Yuan et al., 2021). We find that accompanying sequence-to-sequence log-likelihood ($\log p_{s2s}$) of tokens to hallucination risk helps estimating comprehensive quality more than factual consistency, such as fluency. As in Equation 4, hallucination risk is scaled with a hyperparameter λ, and the log-likelihood of tokens is added to

[1] Appendix Table B.1

[2] We implemented empty input ($\{\}$) as a sequence with only begin and end of the sequence token, namely [BOS], and [EOS]

[3] p_{lm} is not negligible for computing HaRiM (Appendix, Figure A.4).

form HaRiM$^+$.

$$\text{HaRiM}^+ = \frac{1}{L}\sum_{i}^{L}\log(p(y_i|y_{<i};X)) - \lambda * \text{HaRiM} \quad (4)$$

In our experiments, we used $\lambda = 7$, which is a value coherent with the works of Miao et al..[4]

4 Experiments

4.1 Summarization Quality Benchmarks

4.1.1 Factual Consistency Benchmarks

We choose FRANK (Pagnoni et al., 2021), and QAGS annotations (Wang et al., 2020) as benchmarks for assessing the metrics' power to resolve the factuality of article-summary pairs. FRANK and QAGS contain 2246 and 470 pairs, respectively, of article and system-generated summary from CNN-DailyMail (Nallapati et al., 2016), as well as BBC-XSUM (Narayan et al., 2018) corpora. Every pair in the benchmark contains human judgement on factuality. Both benchmarks have similar purpose and annotation format, but differ in annotating environment and aggregation process of the annotations. For FRANK, factual pairs are the intact examples remaining after the annotating errors of each summary introduced by number of annotators, but in QAGS, annotators are directly asked to label each pair if it is factually consistent. We report separate results on each testbed.

In the case of FRANK, the authors recommend measuring partial correlation by considering the confounding variable, the summarization system where summaries are generated from, which can undermine the gaps between metric performances. However, we do not follow this suggestion and conduct experiments with the same setting as others.[5]

4.1.2 Comprehensive Quality Benchmark

SummEval (Fabbri et al., 2021) contains 1600 annotated article-generated summary pairs from 16 summarization systems. The benchmark lets annotators answer about four criteria that a good summary pair should satisfy: coherence, consistency, fluency, and relevance. Each criterion attributes to whether a certain summary is well-organized in structure, factually consistent, grammatically fluent, and containing relevant information regarding the message of the article, respectively. SummEval is comprised of outputs from both abstractive and extractive summarization models which allows dimensional analysis for metrics' performance. We use only the annotations from experts, excluding the ones from turkers, in accordance with the other works' practice using the SummEval for benchmarking (Scialom et al., 2019, 2021; Liu et al., 2022b).[6]

4.2 Measures for Meta-evaluation of Metrics

Measures for describing correlation between two variables are as follows:

- **Kendall's** τ measures how good the metric is ranking the examples (article-summary pairs) in order of human judgement.
- **Spearman's** r assesses how well the relation between the metric and human judgement can be described as monotonic function.
- **Pearson's** ρ measures how linear the metric score is. This may not represent monotonic increment or decrement to the human annotations, but represents proper scaling of the metric; i.e. A metric score should increase linearly according to increment of the judgement score.

All three coefficients range from 0 (independent) to 1 (completely correlated). We report metric-human correlation in τ, and metric-metric correlation with ρ. We find that trends of all three measures move together in our case, and we report τ correlation as the primary measure in our meta-evaluation results in Table 1. Correlations in other measures are reported on Appendix (Table B.3) for further information.

4.3 Metrics

4.3.1 Traditional Metrics

We benchmark traditional N-gram matching baselines; ROUGE-1, 2, L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), sacreBLEU (Riddell et al., 2021) on three benchmarks.[7] For matching-based metrics, we test not only matching to the

[4] λ is determined primarily based on metric correlation to human judgements, but with the consideration of scales of each (Appendix, Figure A.5).

[5] We provide a graphical model representing our claim in Appendix (Figure A.6). Reporting partial correlation to consider the bias introduced by generation system artifacts in the text might help alleviate the vulnerability of a metric, but, in principle, metric does not refer to any other attribute than the text. Thus we decided not to follow the practice of the original benchmark.

[6] In Appendix Figure B.3, We also discuss about reasons why turker annotations are less preferred in discussion section, which supports the arguments from the original authors.

[7] For implementation details, please refer to Appendix C.

reference summaries but also to the article (noted as '_art'), which is reported to benefit metrics assessing factual coverage of the summary (Pagnoni et al., 2021). Additionally, we report some of the relevant statistics; length, and ratio of novel N-gram (Fabbri et al., 2021) in the summary as a metric to compare.

4.3.2 Unsupervised Matching

We also test our metric against the relatively recent matching-based metric based on contextual embedding, BERTScore (Zhang et al., 2019). BERTScore borrows representation power of the pretrained masked language model, BERT (Devlin et al., 2019), to match contextualized embeddings of two texts.

We used roberta-large (Liu et al., 2019) checkpoint provided as default by the package.[8] As done for N-gram metrics, matching toward article is also reported with '_art' notation.

4.3.3 Text Generation Task as an Evaluation

BARTScore (Yuan et al., 2021) reformulated text quality evaluation as a text generation problem. BARTScore depends on the log-likelihood of the fine-tuned BART model to score the quality of the text; averaged log-likelihood of a text is a quality estimation. In our experiments, we test two versions of BARTScore introduced in the original paper. One is BART-large fine-tuned on CNN-DailyMail corpus (Lewis et al., 2020), the other is further fine-tuned to ParaBank2 corpus (Hu et al., 2019) to better capture factual consistency of the article-summary pairs.[9] We also augment BARTScore with hallucination risk to test its correlation toward human judgements. Another objective used as a metric is from CBMI (Zhang et al., 2022), which re-weights negative log-likelihood loss with the conditional bilingual mutual information approximated from token statistics. We flipped the sign of the loss for it to work as higher-better metric.[10]

4.3.4 Question Answering as an Evaluation

Metrics in QA generally require question generation and answering modules that check whether the summary is factually supported by the article. We refer to FEQA (Durmus et al., 2020) and QAGS (Wang et al., 2020) to examine the performance of the QA-based metrics. We benchmark QAGS on two factuality benchmarks, FRANK and QAGS. On QAGS annotations, we re-run QAGS from the original repository[11] to score towards the benchmark. On FRANK, we reused the QAGS and FEQA scores publicly shared on FRANK repository.[12]

4.3.5 Proposed Method: HaRiM$^+$

HaRiM$^+$, our proposed method, exploits summarization model for calculating HaRiM and complement it with log-likelihood, as in Equation 4 to make the final metric score. We use the same summarization model checkpoints as BARTScore as described above for direct comparison: BART-large+cnn (Lewis et al., 2020), and BART-large+cnn+para (Yuan et al., 2021). In the ablation study (Section 5.2), we added another checkpoint, BRIO (Liu et al., 2022a) which also has the same architecture with BART-large.

5 Results

In the followings, we report (1) metric to human judgement correlation in Kendall's τ rank coefficient, and (2) qualitative examples that reveals inductive bias of the hallucination risk (HaRiM$^+$) we proposed. Comparisons with reported values of several other works are attached to Appendix (Table B.1).

5.1 Metric-Human Correlation

Table 1 shows the metric to human judgement (segment-level)[13] correlation. Proposed HaRiM$^+$ records highest Kendall's τ in most criteria of *CNN/DailyMail* based benchmarks. To thoroughly show the significance test result, we attach permutation test matrix on Figure A.1 in Appendix. Because HaRiM$^+$ and BARTScore shares the same summarization model, both metrics with respective models show similar scoring patterns. HaRiM$^+$ records mostly highest correlation toward human judgements except several settings (XSUM, and SummEval-Relevance). For SummEval relevance score benchmark, BERTScore P_art outperforms the HaRiM$^+$ (BART_large + cnn) by 0.024 points, which indicates BERTScore P_art is 1.2%p better at ranking hallucinated results. In FRANK-XSUM benchmark, despite using a summarization model

[8] https://github.com/Tiiiger/bert_score

[9] Model checkpoints for BARTScore are from https://huggingface.co/facebook/bart-large,https://github.com/neulab/BARTScore.

[10] For detailed information of implementation, refer to Appendix C.5.

[11] https://github.com/W4ngatang/qags

[12] https://github.com/artidoro/frank

[13] system level correlation reported in Table B.2

trained on *CNN/DailyMail*, HaRiM$^+$ records high score ($\tau = 0.141$ compared to $\tau = 0.151$ of BERTScore P). On FRANK-CNNDM, we perform a permutation test to confirm that HaRiM$^+$ outperforms the others with the confidence $>.95$ which is attached to the Appendix (Table A.2) for space issue.[14] Overall, HaRiM$^+$ records robust performances in ranking the summary pairs according to the human judgement for CNN-DailyMail corpus examples which the core model is trained to, while it also scored high on XSUM corpus.

5.2 Ablation Study: Effect of Accompanying Log-likelihood

We conduct ablation study on HaRiM$^+$ varying the model checkpoints. HaRiM$^+$ is compared to each term used in single: log-likelihood, and the regularization term only (HaRiM). Table 2 shows the results for the average scores across all four SummEval criteria; the table indicates that accompanied use of log-likelihood with HaRiM (that is, HaRiM$^+$ helped complementing the metric performance.

5.3 Qualitative Analysis: Detecting Hallucinations

We test the HaRiM$^+$ (BART-large+cnn) under hallucination detecting scenario to provide hint for how HaRiM$^+$ behaves in various summary outputs. In Table 3, we randomly pick an article from *CNNDailyMail* split of the FRANK benchmark and prepare several summaries. We collected the following five summaries to pair with the article: (1) reference target summary, (2) summary generated from BART-large+cnn (Self-generation), (3) unfactual summary of summarization model (displayed example is generated by RNN-S2S (Sutskever et al., 2014)), (4) reference summary permutation with wrong subject, which contains wrongly-injected subject entity from the source article, and (5) a negated reference summary.

As shown in Table 3, we align the summary with HaRiM$^+$ (BART-large+cnn) score and its score gain compared to the reference summary score. HaRiM$^+$ metric ranks the summaries in order of self-generated>reference>permuted references>wrong generation. We attribute the HaRiM$^+$ metric's preference toward self-generation to inductive bias: both the self-generation model and HaRiM$^+$ evaluation model are the exact twins. To roughly put, the self-generation model works as an oracle summary generator for the metric. The inductive bias of HaRiM$^+$ metric will be discussed further with quantitative evidence in Section 6.1. The trend of ranking factual human-written summaries over unfactual summaries, which includes permutated references, are observed constantly throughout the *CNNDailyMail* corpus examples. We provide several more examples in Appendix (Table B.6, B.7, B.8, B.9, and B.10).

6 Discussion

6.1 Inductive Bias

As mentioned in qualitative analysis, the metric has inductive bias of preferences toward summaries generated by abstractive summarization systems. Proposed HaRiM$^+$ prefers self-generated summary (i.e. summary generated by the same summarization model the scorer depending on) to human written references. Another hint for this bias could be found when we dissect the SummEval benchmark results into abstractive and extractive summary splits. In Table 4, not only log-likelihood but also regularization term, HaRiM, both prefer outputs from abstractive system. As summary text becomes similar to the evaluating summarization model's likely output, generation-based metrics (including HaRiM$^+$) become more generous at scoring. In other word, how bad the assessed summary would not be a problem if the summarizer used for evaluation resembles the system which wrote the summary being assessed. In this context, using the model trained on too noisy dataset, without proper regularization would result in unreliable evaluation. Figure 2 shows how noisy summarization models could be trained under-regularized; most of the output summary trained on *XSUM* with MLE strategy contain errors. Therefore, we decide not to exploit summarization model fine-tuned on *XSUM* even if it could result in better correlation on FRANK/QAGS-XSUM splits.

6.2 Metric Performance of HaRiM$^+$ in Machine Translation

We also tested our metric, HaRiM$^+$, on WMT20 metrics task (Mathur et al., 2020) to see whether HaRiM$^+$ works in the machine translation domain (Table 5). WMT20 DA annotation contains machine translation pairs of language pairs accompa-

[14] Several notable observations in metric-metric correlation had to be pushed back to Appendix (e.g. NovelNgram highly correlates (>.6) to BERTScore_art, and HaRiM$^+$, but HaRiM$^+$ and BERTScore_art are not).

Kendall's τ	CNNDM						XSUM	
	FRANK	QAGS	SummEval				FRANK	QAGS
Metrics	**Factuality**	**Factuality**	**Con**	**Coh**	**Flu**	**Rel**	**Factuality**	**Factuality**
N-gram-matching								
ROUGE 1	0.182	-0.052	0.105	0.123	0.062	0.209	0.125	0.110
ROUGE 2	0.135	-0.107	0.101	0.097	0.048	0.153	0.128	0.097
ROUGE L	0.141	-0.072	0.091	0.113	0.061	0.164	0.117	0.090
METEOR	0.198	0.053	0.125	0.116	0.070	0.223	0.121	0.115
sacreBLEU	0.136	-0.085	0.080	0.167	0.088	0.131	0.113	0.012
ROUGE 1_art	0.185	0.243	0.111	0.036	0.058	0.127	-0.003	-0.074
ROUGE 2_art	0.249	0.315	0.195	0.072	0.119	0.165	0.027	0.069
ROUGE L_art	0.225	0.305	0.203	0.097	0.123	0.050	0.010	-0.019
METEOR_art	0.174	0.234	0.112	0.009	0.071	0.091	0.004	-0.052
sacreBLEU_art	0.153	0.245	0.091	0.042	0.035		-0.038	-0.139
N-gram stats								
NovelNgram_4	0.275	0.392	0.221	0.203	0.173	0.205	0.017	0.056
NovelNgram_3	0.273	0.370	0.218	0.208	0.171	0.208	0.064	0.080
NovelNgram_2	0.259	0.327	0.199	0.209	0.150	0.207	0.053	0.129
NovelNgram_1	0.219	0.201	0.090	0.190	0.068	0.173	0.091	0.120
Length (no. tokens)	0.187	0.185	0.078	0.033	0.000	0.000	-0.111	-0.132
Contextual Embedding								
BERTScore P	0.168	-0.067	0.041	0.229	0.097	0.192	**0.151**	0.016
BERTScore R	0.250	0.017	0.125	0.241	0.097	0.299	0.107	0.058
BERTScore F1	0.232	-0.029	0.079	0.267	0.111	0.267	0.142	0.036
BERTScore P_art	0.301	0.331	0.266	0.308	0.236	**0.308**	0.038	-0.039
BERTScore R_art	0.360	0.365	0.141	0.153	0.112	0.234	0.144	-0.022
BERTScore F1_art	0.358	0.365	0.230	0.256	0.192	0.307	0.111	-0.040
Neural entailment								
FactCC (Kryscinski et al., 2020)	0.376						0.071	
Dep Entail (Goyal and Durrett, 2020)	0.342						0.092	
Q&A based								
FEQA (Durmus et al., 2020)	-0.008						0.006	
QAGS (Wang et al., 2020)	0.206	0.274					-0.006	**0.153**
QAEval-F1 (Deutsch et al., 2021a)						0.220*	-0.006	**0.153**
Text Generation based								
CBMI (BART_base + cnn)	0.058	0.026	0.152	-0.029	0.023	0.208	-0.077	-0.041
BARTScore (BART_large+cnn) (Yuan et al., 2021)	0.413	0.470	0.197	0.310	0.181	0.263	0.137	0.072
BARTScore (BART_large+cnn+para) (Yuan et al., 2021)	0.392	0.416	0.259	0.301	0.238	0.278	0.145	0.031
Proposed								
HaRiM$^+$ (BART_large + cnn)	**0.424**	**0.478**	0.251	**0.315**	0.210	0.284	0.136	0.076
HaRiM$^+$ (BART_large + cnn + para)	0.399	0.401	**0.281**	0.293	**0.245**	0.282	0.141	0.028

Table 1: Metric-to-human judgement correlation (segment level) reported in Kendall's τ. **Bold**-face values are the largest correlating metrics, underlined are second-large values amongst the metrics. HaRiM$^+$ outperforms others in most criteria. SummEval's quality criteria; consistency, coherence, fluency, and relevance are abbreviated as Con, Coh, Flu, and Rel respectively. We provide permutation test result and results in Spearman's r and Pearson's ρ in Appendix (Figure A.1, Table B.3). In Table B.1, we also provide comparisons to reported values that could not be directly presented above. *:correlation value taken from (Deutsch et al., 2021a)

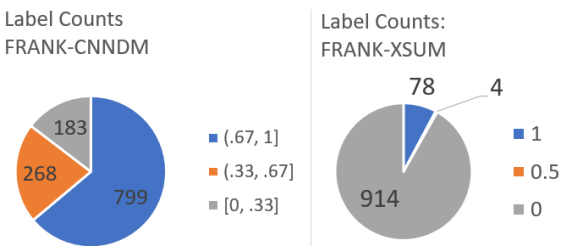

Figure 2: Factuality label counts from FRANK benchmark. Legend shows the value of factuality annotation, varying from 0 (unfactual) to 1 (factual). The factuality labels for XSUM corpus are almost binary.

Checkpoints	Log-likelihood	HaRiM	HaRiM$^+$
BART-large + cnn	0.238	**0.279**	0.265
BART-large + cnn + para	0.269	0.256	**0.275**
BRIO (Liu et al., 2022a)	0.262	0.252	**0.265**

Table 2: Effect of accompanied use of log-likelihood and regularization term HaRiM

nied with human judgements of quality. We find that there is little improvement in correlation to human annotation in several language pairs, but it is not significant in average of all language pairs. In case of WMT20 metrics task, performance of the generation-based metrics seems to rely heavily on generation model checkpoints and its train corpus

Source Article
Spain's 2-0 defeat by Holland on Tuesday brought back bitter memories of their disastrous 2014 World Cup, but coach Vicente del Bosque will not be too worried about a third straight friendly defeat, insists Gerard Pique. Holland, whose 5-1 drubbing of Spain in the group stage in Brazil last year marked the end of the Iberian nation's six-year domination of the world game, scored two early goals at the Amsterdam Arena and held on against some determined Spain pressure in the second half for a 2-0 success. (...) Stefan de Vrij (right) headed Holland in front against Spain at the Amsterdam Arena on Tuesday Gerard Pique (left) could do nothing to stop Davy Klaassen doubling the Dutch advantage Malaga forward Juanmi and Sevilla midfielder Vitolo became the 55th and 56th players to debut under Del Bosque, (...) 'The national team's state of health is good,' centre back Gerard Pique told reporters. 'We are in a process where players are coming into the team and gathering experience,' added the Barcelona defender. 'We are second in qualifying (for Euro 2016) and these friendly games are for experimenting. (...)

Model	Summary	HaRiM$^+$ Score ↑	Score Gain ↑
Reference	holland beat spain 2-0 at the amsterdam arena on tuesday night . stefan de vrij and davy klaassen scored goals for holland . defeat recalls horror 5-1 defeat by holland at the world cup . vicente del bosque used game to give younger spain players a chance .	1.6247	-
Self-generation (BART-large+cnn)	holland's 5-1 drubbing of spain last year marked the end of the iberian nation's six-year domination of the world game. spain's 2-0 defeat by holland on tuesday brought back bitter memories of their disastrous 2014 world cup, but coach vicente del bosque will not be too worried about a third straight friendly victory. 'the national team's state of health is good,' says defender gerard pique	3.7446	+2.1200
RNN-S2S (Factuality=0.0)	holland beat spain 2-0 in the group stage in brazil on tuesday night . del bosque will be hoping to find the right mix of players to the world cup . gerard pique could make the right mix of players to the tournament .	0.1173	-1.5074
Reference (w/ wrong subject)	del bosque beat spain 2-0 at the amsterdam arena on tuesday night . stefan de vrij and davy klaassen scored goals for holland . defeat recalls horror 5-1 defeat by holland at the world cup . vicente del bosque used game to give younger spain players a chance .	1.3229	-0.3017
Reference (w/ negation)	holland could not beat spain 2-0 at the amsterdam arena on tuesday night . stefan de vrij and davy klaassen scored goals for holland . defeat recalls horror 5-1 defeat by holland at the world cup . vicente del bosque used game to give younger spain players a chance .	1.4132	-0.2115

Table 3: Testing HaRiM$^+$ metric under hallucination detecting scenario. Part of the source article, which is irrelevant to the summaries are omitted for clarity. The words highlighted red are hallucinated information deliberately injected to the reference.

		abstractive	extractive	Δ
BART -Large	log-likelihood	0.266	0.160	0.106
	HaRiM	0.303	0.174	0.129
	HaRiM$^+$	0.293	0.168	0.125
BRIO	log-likelihood	0.308	0.143	0.165
	HaRiM	0.295	0.117	0.177
	HaRiM$^+$	0.311	0.137	0.174
BART -Score	log-likelihood	0.296	0.168	0.128
	HaRiM	0.280	0.150	0.130
	HaRiM$^+$	0.303	0.166	0.137
Average		0.295	0.154	0.141

Table 4: Averaged τ correlation on SummEval. Δ indicates difference of τ coefficients measured toward abstractive and extractive summaries.

	sys(ρ)		seg(τ)*	
	all	**all-out**	**all**	**all-out**
(1) BART-large+cnn+para→MBART50_m2m				
Log-likelihood	-0.001	-0.005	-0.020	-0.024
HaRiM$^+$	0.002	0.000	-0.016	-0.020
(2) Log-likelihood→HaRiM$^+$				
BART-large+cnn+para	+0.001	0	0	-0.001
PRISM(m39v1)	0	0	0	+0.001
MBART50_m2m	0	+0.002	+0.001	+0.002

Table 5: Change of generation-based metric performance according to (1) model weight change (2) applying HaRiM$^+$. All results are averaged over language pairs from data supported by each model (i.e. BART-large+cnn+para averages the results of only 'to English' language pairs). Note that τ we use here is WMT-variant suggested in (Barrault et al., 2021). For fair comparison, in (1), only 'to English' pairs are used. For MBART (Liu et al., 2020) we used mbart50-many-to-many model, for PRISM (Thompson and Post, 2020), we used m39v1 model.

distribution rather than the hallucination risk consideration. As WMT metrics task has a broad range of dimensions to explore, we leave this as a future remark for generation-based evaluation metrics and text generation models.

7 Conclusion

In this study, we propose HaRiM$^+$ as a new summarization metric, which exploits the power of the summarization model for evaluation accompanied with the hallucination risk into consideration. For

evaluating summaries, HaRiM$^+$ only requires the summarization model without further training, additional module, or complicated pipelines. Our method further demonstrates the merit of using summarization models not only for summary generation but also for evaluation. Throughout the quantitative and qualitative analyses, we show that the HaRiM$^+$ metric correlates well to human judgment in comprehensive aspects with robust performance, demonstrated with qualitative examples. We also explored the inductive bias of the model, which emphasizes the importance of training noisy-robust summarization-generation models for evaluation use. We leave the potential extension of the metric to another generation task, such as machine translation, as a future remark.

References

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

Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Alexander 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.

Yanjun Gao, Chen Sun, and Rebecca J. Passonneau. 2019. Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418, Hong Kong, China. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Aaron Harnly, Ani Nenkova, Rebecca Passonneau, and Owen Rambow. 2005. Automation of summary evaluation by the pyramid method. In *Recent Advances in Natural Language Processing (RANLP)*, pages 226–232.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia. Association for Computational Linguistics.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, 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.

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

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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

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

Yu Lu Liu, Rachel Bawden, Thomas Scaliom, Benoît Sagot, and Jackie Chi Kit Cheung. 2022b. Maskeval: Weighted mlm-based evaluation for text summarization and simplification. *arXiv preprint arXiv:2205.12394*.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

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

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1080–1089, Online. Association for Computational Linguistics.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3456–3468, Online. 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.

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.

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

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rebecca J. Passonneau, Emily Chen, Weiwei Guo, and Dolores Perin. 2013. Automated pyramid scoring of summaries using distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–147, Sofia, Bulgaria. Association for Computational Linguistics.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

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(140):1–67.

Allen Riddell, Haining Wang, and Patrick Juola. 2021. A call for clarity in contemporary authorship attribution evaluation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1174–1179, Held Online. INCOMA Ltd.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. 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.

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. 2021. Conditional adversarial networks for multi-domain text classification. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 16–27, Kyiv, Ukraine. Association for Computational Linguistics.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Songming Zhang, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, Jian Liu, and Jie Zhou. 2022. Conditional bilingual mutual information based adaptive training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2377–2389, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

A Additional Results

A.1 Comparison of HaRiM$^+$ Performance to Reported Values

We separately represent the meta-evaluation results compared to reported metrics' benchmark scores in Table B.1. Mostly the reported values are using r and ρ to estimate metric performance, which does not fit into our selection of primary means of measure (τ). Reason for avoiding the use of ρ is simple: ρ does not guarantee monotonic relation between correlated variables, rather it means linearity, and we found τ to be more interpretable measure for ranking the quality of article-summary pairs.

A.2 System-level Metric-Human Correlations on SummEval

In Table B.2, we report system-level correlation of metric scores on SummEval benchmark, which contains total 16 systems. To 100 articles, 16 systems (12 abstractive, 4 extractive) present their summary generation.

A.3 Metric-Human Correlations in Spearman's r and Pearson's ρ

In Table B.3, we provide benchmark results with Spearman's rank coefficient (r), and Pearson's ρ. As mentioned earlier, for our set of metric scores, three correlation measures orders almost the same with each other while it is not guaranteed in general.

A.4 Significance by Randomization Test

With randomization test in Figure A.1, we can compute the confidence of the difference being coincidant by chance or significant with certain confidence. We follow the practice of (Deutsch et al., 2021b), PERM-INPUT, as our correlation benchmarking only covers summary-level metric score alignment to human judgement. We provide randomization test results for every pair of metrics on metric-human correlation on FRANK benchmark, which provides the largest number of metrics are available. HaRiM$^+$ largely outperforms the others.

A.5 Metric-Metric Correlation

In Figure A.2 and A.3, We provide metric-metric correlation with Pearsonś ρ which might hint the similarity between metric behaviors. We highligted several notable trend similarity of the metrics with the red boxes on Figure A.2 according to the following criteria: ρ rounds to .7 or larger, while not a clearly relevant metric (around the diagonal).

Observation shows that text-generation-based metrics correlates well with NovelNgram variants and BERTScore_art (P, F1, not R) while not with ROUGE. BERTScore behavior differs quite much when applied to article or reference. BERTScore measured with reference text resembles behavior of ROUGE scores while they turns more similar to NovelNgrams and text-generation-based metrics (HaRiM$^+$, and BARTScore) for BERTScore-P (BS P_art). CBMI, is the most resemblant metric to length of the summary text (L) which records 0.72 in ρ.

A.6 SummEval Separate Results: Abstractive/Extractive System Outputs

In Figure B.5, we provide benchmark results (τ correlation) toward abstractive and extractive summary outputs in separate. As discussed in the Section 6.1, HaRiM$^+$ correlates better on abstractive system outputs.

A.7 More of Qualitative Examples

We present several more qualitative examples in Table B.6, B.7, B.8, B.9, and B.10. Those five examples are from FRANK benchmark, three are showcasing hallucinated outputs (Factuality=0) and following two are for factual outputs (Factuality=1).

B Analyses

B.1 HaRiM variations tested on FRANK

In Table B.4, we show our heuristic trials to aggregate $\Delta = p_{s2s} - p_{lm}$ to make the hallucination risk (HaRiM) better correlate to the human judgements in FRANK benchmark. We found the original form, denoted as *linear*, works stable than the others. Applying other function-form (log or exponential) than linear for $\Delta(= p_{s2s} - plm)$ was not effective. Also for aggregating token level scores, we tried applying `tfidf` and `idf`, which turned out doing nothing than worsening the correlation as similarly top/bot 5 average do. Entropy-based scores are also tested but found ineffective.

B.2 Effect of variables to HaRiM

We show fine-grained effect of each variables (e.g. p_{lm}, p_{s2s}, Δ) to HaRiM. Figure 1 shows article-summary pair as a datapoint in the plot, here we show each token of the decoded output as a datapoint. Replacing p_{lm} with empty-sourced de-

coder inference looks fair even in token-level plot (HaRiM did not change drastically). HaRiM seems quite dependent on p_{s2s}, but as we reported earlier in the main body of this paper (benchmark results), use of p_{lm} quite helps benefits HaRiM$^+$ a lot.

B.3 Why should not the performance on FRANK benchmark reported with partial correlation

The correlation value reported on the Table1, column FRANK shows correlation to human judgements, not considering partial correlation as suggested in (Pagnoni et al., 2021). A metric, or a scorer for the text-quality measurement does not refer to the system which wrote the text while the partial correlation suggested by Pagnoni et al. considers this as a confounding variable that hinders precise meta-evaluation of the metrics. In Figure A.6, we represent our claim that the generation system should not be taken into account for metric meta-evaluation with two graphical models. The graph A shows the view of Pagnoni et al., which considers generation system (i.e. summarization model), into account while the other graph (B) shows ours. Metric score, M, and human judgement, H, are both grounded by the text, which blocks the effect of generation system, S, in the graphical model; which means considering S for measuring the correlation betweeen M and H is at best doubtful for precise meta-evaluation.

B.4 SummEval: Why Experts' Annotations not Turkers'?

In Figure A.7, and A.8, we plotted averaged experts' annotations over annotators and 4 aspects of quality (i.e. consistency, cohenrence, fluency, relevance), versus turkers' counterpart of those. Turkers' judgement of quality in average look irrelevant to correspondings of experts. As mentioned in (Fabbri et al., 2021), expert annotators are re-instructred after the first round of annotation, which resulted improved inter-annotator-agreement. Thus, trusting in annotations from experts but not for crowdworkers of SummEval is plausible as other works done on SummEval benchmark annotation set.

C Implementation Details

C.1 QAGS

QAGS scorer: We used original code from the author (`https://github.com/W4ngatang/qags`) except its missing part which provide functions for matching the generated answer with GT, in SQuAD style.

Aggregating Annotations: "Yes" are considered 1 and "no" considered 0 (coherent to the sign of the FRANK benchmark annotations) to finally obtain averaged factuality label we used. Annotations are also from the original repository.

C.2 BERTScore

We used `BERTScore==0.3.11` (`https://github.com/Tiiiger/bert_score`) which defaults to RoBERTa-large weight for text.

C.3 N-gram Metrics

For traditional N-gram-based metrics, we used huggingface's `datasets.load_metric()` wrapper to load SacreBLEU, METEOR, and ROUGE. Codebase of each metric is as follow:

- SacreBLEU: `sacreBLEU==2.1.0` from the repository (`https://github.com/mjpost/sacrebleu`).
- METEOR: `nltk.translate.meteor_score` from `NLTK=3.6.4`.
- ROUGE: We used `datasets.load_metric('rouge')` which uses `https://github.com/google-research/google-research/tree/master/rouge` as its codebase.

C.4 Novel Ngram

Equation 5 describes our computation of Novel-Ngram, which does not consider duplication of the tokens. Minus sign is applied to use it as a higher-is-better score.

$$\mathrm{NN_i} = -\frac{\mathrm{len(set(Ngram_i^{output}) - set(Ngram_i^{article}))}}{\mathrm{len(set(Ngram_i^{article}))}} \tag{5}$$

C.5 CBMI

Original implementation of conditional bilingual mutual information (CBMI) proposed by Zhang et al. uses minibatch statistics for nomalization. Instead we take whole examples of FRANK benchmark to compute the CBMI statistics.

C.6 List of Reused Metric Scores from FRANK repository

We measured all the other metric scores on all benchmarks other than specified below.

- FactCC (Kryscinski et al., 2020)
- Dependency Arc Entailment (Dep Entail) (Goyal and Durrett, 2020)
- FEQA (Durmus et al., 2020)
- QAGS on FRANK benchmark (Wang et al., 2020; Pagnoni et al., 2021)
  (on QAGS annotation set, we scored with re-implemented scorer)

C.7 Score Scales: HaRiM$^+$, HaRiM, and Log-likelihood

In Figure A.5, we visualize score scales of proposed HaRiM$^+$, HaRiM, and log-likelihood varying summarization model checkpoints. We considered scale of each HaRiM and loglikelihood to decide the mixing coefficient λ (searched over 0.1, 1, 5, 7, 8, 10, 20 and finally chose 7 to use).

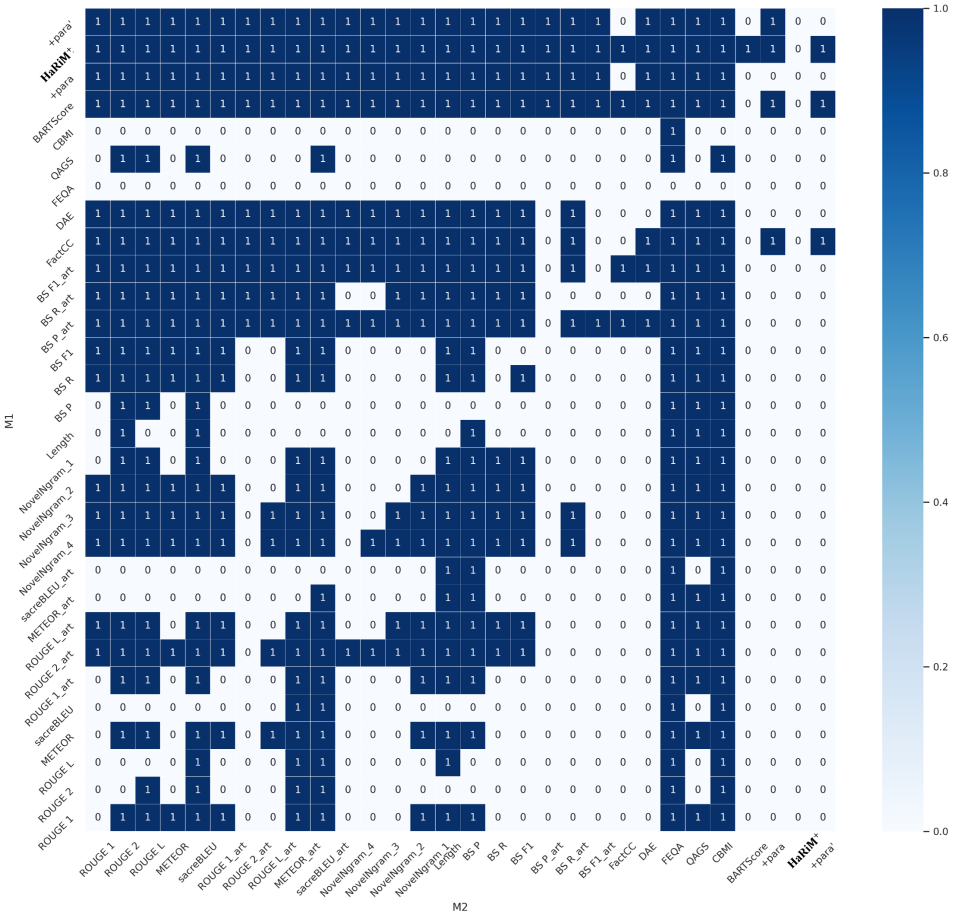

Figure A.1: Permutation test done for metric scores on FRANK-CNN/DM. 1 (filled grid) represents significant difference in metric performance, 0 represents negligible difference with confidence >=.95 ($p <= 0.05$), i.e. HaRiM is significantly more correlated to human judgements than all the other metrics except itself with a confidence of $>=$95%.

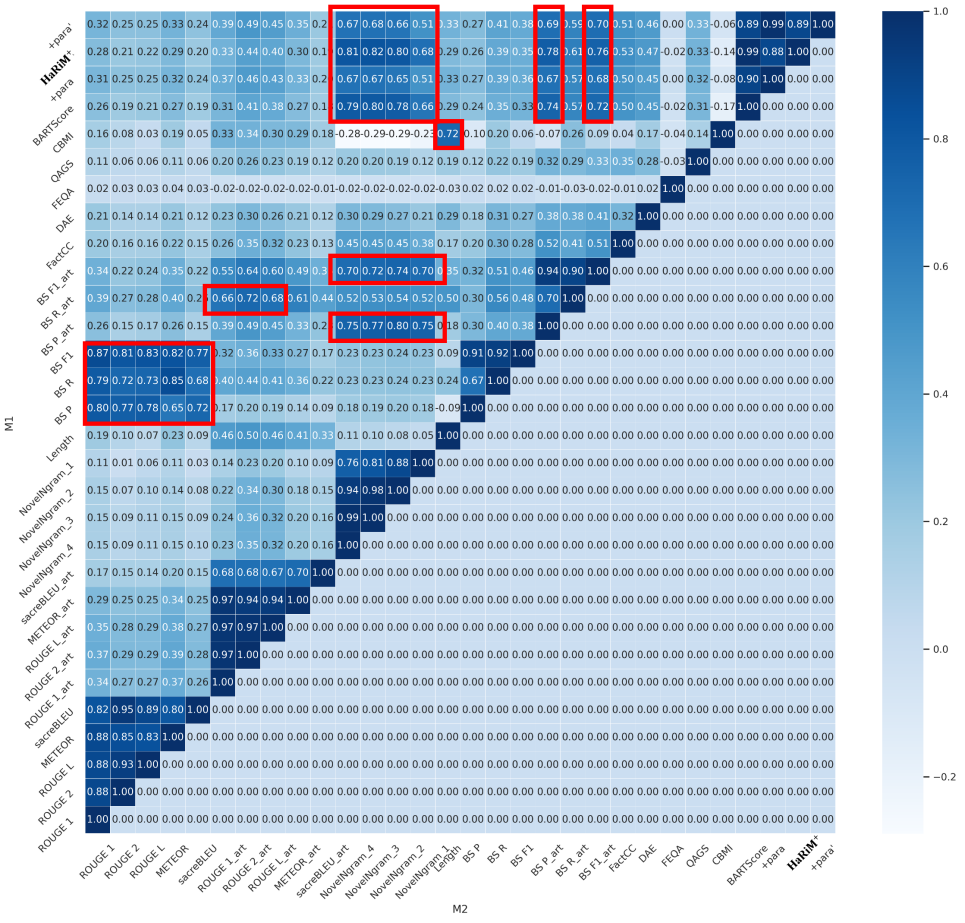

Figure A.2: Pearson's ρ correlation between metric scores on FRANK-CNN/DM split. The highter the correlation, the similar the metric behavior becomes. Red boxes highlights notable observation which is unexpected behavioral similarity between metrics.

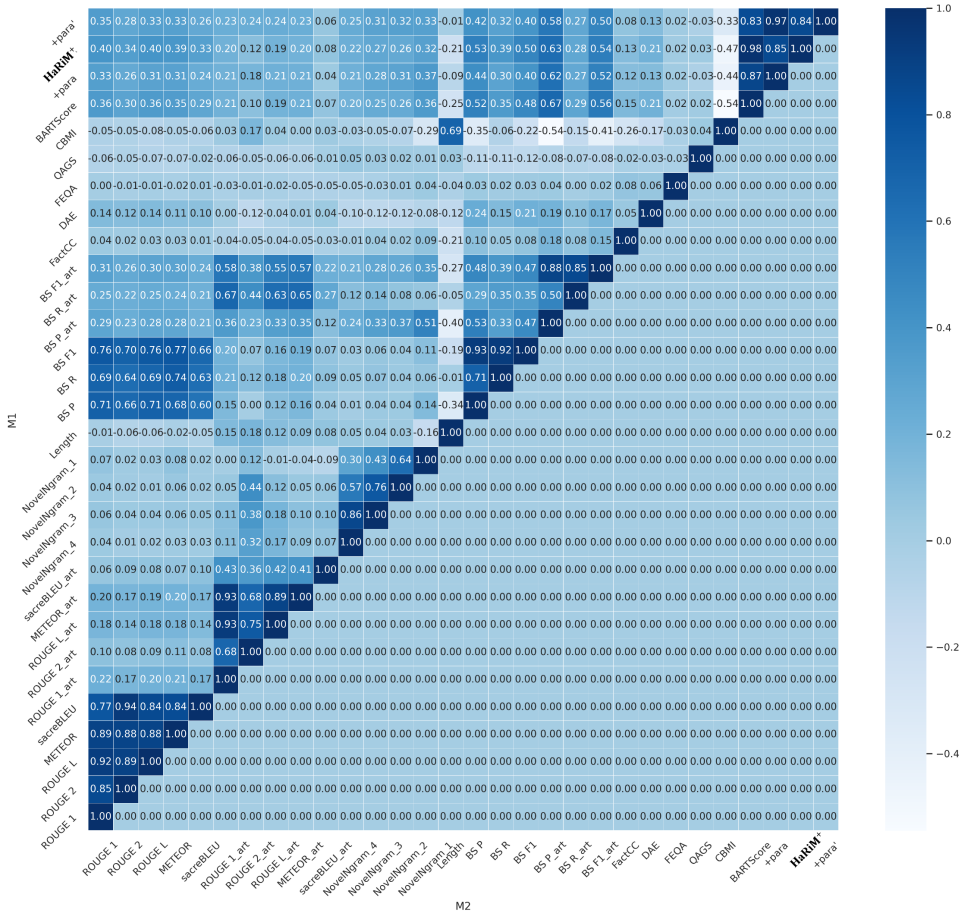

Figure A.3: Pearson's ρ correlation between metric scores on FRANK-BBC/XSUM split. The highter the correlation, the similar the metric behavior becomes.

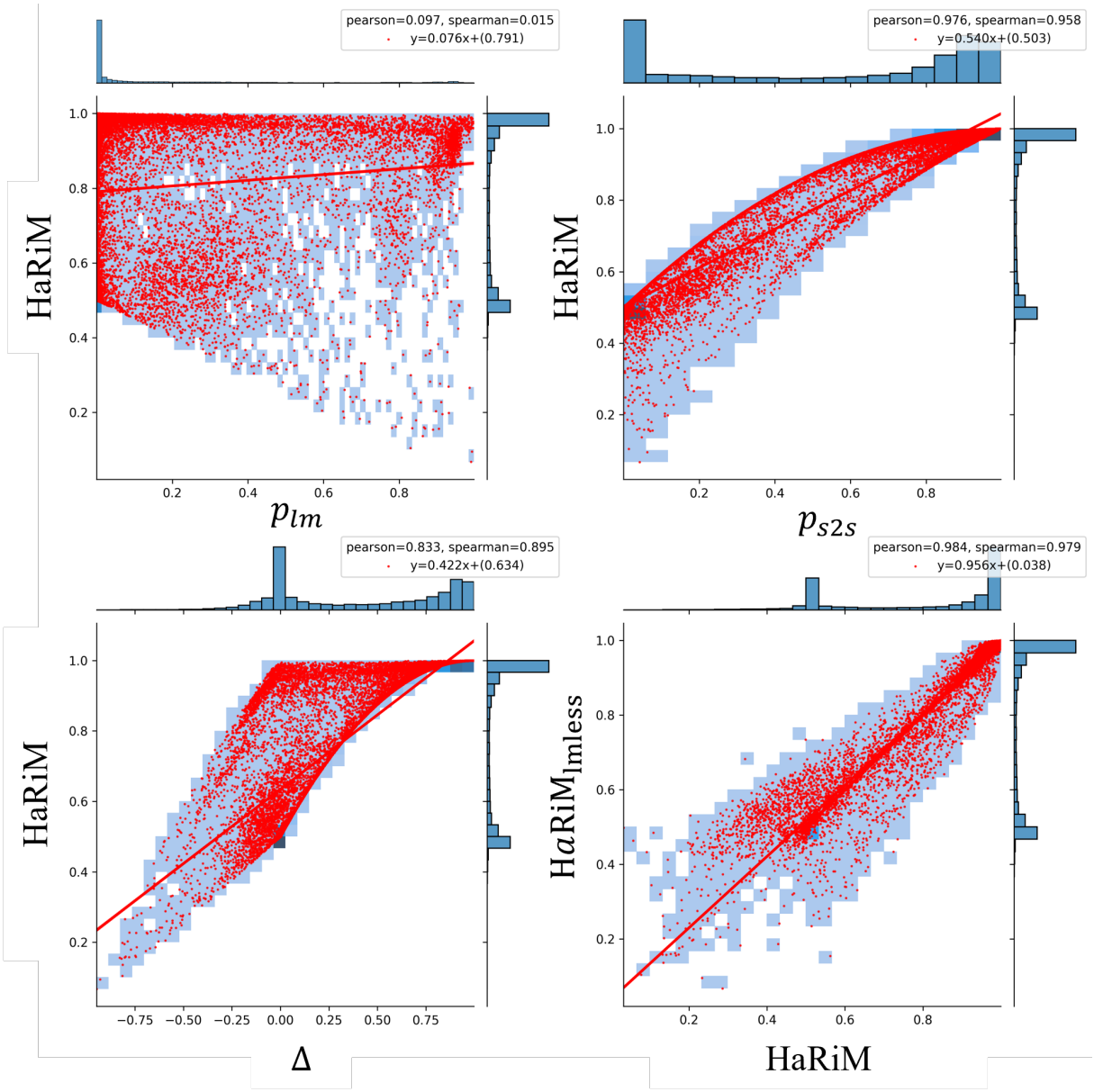

Figure A.4: Effect of each variable to HaRiM. Δ represents $p_{s2s} - p_{lm}$. The last figure at the righter down shows the effect of replacing auxiliary LM probability with empty-sourced decoder inference ($HaRiM_{lmless}$). Figure 1 shows article-summary pair as a datapoint in the plot, here we show each token of the decoded output as a datapoint.

	QAGS-CNNDM		QAGS-XSUM		SummEval (1200 outputs)	
	r	ρ	r	ρ	r	ρ
QAGS	0.382	0.466	0.203	0.217		
FFCI_BERTScore*	0.485	0.486	0.200	0.190	0.285	0.308
QuestEval_F1*	0.492	0.445	0.007	0.010	0.370	0.339
CoCo_span*	0.573	0.501	0.187	0.187	**0.436**	0.410
CoCo_sent*	0.588	0.523	**0.241**	**0.227**	0.420	0.390
HaRiM$^+$ (BART-large+cnn+para)	0.530	0.610			0.405	0.430
HaRiM$^+$ (BART-large+cnn)	**0.620**	**0.679**			0.392	0.415
HaRiM$^+$ (BRIO)	0.514	0.569			0.417	**0.443**

Table B.1: Metric correlation to human judgements on SummEval-abstractive (1200 out of 1600 total examples) QAGS annotation set in Pearson's ρ and Spearman r. * notes that the values are copied from each paper (Xie et al., 2021).

	SummEval (system-level correlation, 16 systems)											
	consistency			coherence			fluency			relevance		
Metrics	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r
n-gram-matching												
ROUGE 1	0.500	0.662	0.688	0.267	0.063	0.459	0.450	0.554	0.635	0.500	0.550	0.682
ROUGE 2	0.600	0.653	0.765	0.233	0.085	0.338	0.483	0.542	0.676	0.433	0.561	0.626
ROUGE L	0.283	0.697	0.385	0.383	0.204	0.506	0.467	0.624	0.600	0.517	0.600	0.712
METEOR	0.550	0.559	0.703	0.017	0.044	0.026	0.267	0.449	0.385	0.250	0.438	0.312
sacreBLEU	-0.050	0.175	-0.118	0.383	0.493	0.529	0.233	0.233	0.318	0.283	0.462	0.418
ROUGE 1_art	0.467	0.467	0.626	0.000	0.028	-0.068	0.217	0.375	0.288	0.200	0.324	0.174
ROUGE 2_art	0.500	0.599	0.688	0.067	0.072	-0.026	0.283	0.515	0.329	0.267	0.370	0.212
ROUGE L_art	0.550	0.618	0.726	0.117	0.164	0.018	0.300	0.541	0.362	0.317	0.421	0.265
METEOR_art	0.467	0.513	0.621	0.000	0.082	-0.021	0.250	0.430	0.335	0.233	0.397	0.226
sacreBLEU_art	0.450	0.287	0.621	0.083	0.299	0.176	0.200	0.277	0.318	0.183	0.351	0.209
N-gram stats												
NovelNgram_4	0.400	0.623	0.553	0.300	0.704	0.435	0.450	0.691	0.606	0.367	0.664	0.506
NovelNgram_3	0.367	0.590	0.512	0.333	0.657	0.453	0.417	0.649	0.594	0.367	0.631	0.506
NovelNgram_2	0.300	0.464	0.444	0.367	0.615	0.524	0.417	0.522	0.576	0.400	0.570	0.541
NovelNgram_1	-0.017	0.016	0.006	0.417	0.456	0.529	0.167	0.091	0.241	0.183	0.276	0.244
Length (no. tokens)	0.417	0.348	0.571	-0.050	-0.009	-0.112	0.200	0.262	0.268	0.183	0.239	0.156
Contextual Embedding												
BERTScore P	-0.233	-0.254	-0.341	0.300	0.457	0.406	0.017	-0.122	0.047	0.067	0.126	0.150
BERTScore R	0.617	0.459	0.809	0.550	0.671	0.697	0.600	0.486	0.806	0.617	0.749	0.797
BERTScore F1	0.017	-0.039	0.021	0.550	0.623	0.715	0.333	0.083	0.432	0.417	0.373	0.497
BERTScore P_art	0.583	0.654	0.809	0.450	0.715	0.559	0.500	0.691	0.662	0.550	0.714	0.635
BERTScore R_art	**0.750**	0.623	**0.903**	0.317	0.441	0.453	0.567	0.589	0.756	0.517	0.653	0.676
BERTScore F1_art	0.683	0.680	0.868	0.417	0.623	0.559	0.600	0.684	0.753	0.583	0.727	0.691
Text Generation based												
CBMI (BART_base + cnn)*	0.433	0.483	0.632	-0.033	-0.119	-0.132	0.217	0.384	0.238	0.200	0.185	0.132
BARTScore (BART-large + cnn)**	0.183	0.301	0.259	0.717	0.812	0.871	0.467	0.423	0.559	0.550	0.592	0.621
BARTScore (BART-large + cnn + para)**	0.283	0.577	0.424	0.650	**0.891**	0.809	0.567	0.687	0.735	0.617	0.783	0.750
Proposed												
HaRiM+ (BART_large + cnn)	0.250	0.492	0.368	**0.817**	0.835	**0.926**	0.500	0.593	0.679	0.650	0.721	0.756
HaRiM+ (BART_large + cnn + para)	0.383	**0.701**	0.562	0.617	0.860	0.762	**0.667**	**0.790**	**0.859**	**0.717**	**0.851**	**0.859**

Table B.2: System-level correlation on SummEval, total 16 systems (12 abstractive, 4 extractive). **Bold**face numbers represent the best and underlined are the second-best. We omit abstractive-systems-only result as its trend is similar to above.

	CNNDM												XSUM			
	FRANK		QAGS		SummEval								FRANK		QAGS	
	Factuality		Factuality		con		coh		flu		rel		Factuality		Factuality	
Metrics	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
n-gram-matching																
ROUGE 1	0.239	0.254	-0.072	-0.013	0.167	0.133	0.181	0.175	0.136	0.080	0.323	0.289	0.153	0.179	0.148	0.163
ROUGE 2	0.178	0.181	-0.151	-0.019	0.147	0.128	0.131	0.138	0.087	0.062	0.240	0.234	0.154	0.186	0.134	0.145
ROUGE L	0.186	0.194	-0.100	-0.042	0.142	0.115	0.155	0.160	0.110	0.079	0.248	0.231	0.144	0.182	0.121	0.117
METEOR	0.260	0.268	0.074	0.050	0.173	0.158	0.168	0.165	0.114	0.091	0.360	0.312	0.148	0.165	0.156	0.157
sacreBLEU	0.179	0.169	-0.116	-0.063	0.117	0.102	0.250	0.238	0.139	0.113	0.290	0.290	0.139	0.156	0.016	0.036
ROUGE 1_art	0.244	0.255	0.336	0.355	0.137	0.142	0.074	0.049	0.087	0.075	0.209	0.179	-0.004	-0.017	-0.103	-0.065
ROUGE 2_art	0.327	0.331	0.427	0.475	0.252	0.247	0.123	0.099	0.188	0.154	0.245	0.215	0.033	0.012	0.091	0.107
ROUGE L_art	0.296	0.297	0.411	0.462	0.242	0.258	0.155	0.133	0.177	0.159	0.252	0.230	0.012	0.000	-0.024	0.014
METEOR_art	0.229	0.230	0.324	0.277	0.122	0.143	0.053	0.011	0.093	0.091	0.150	0.129	0.005	-0.005	-0.071	-0.015
sacreBLEU_art	0.202	0.093	0.337	0.180	0.073	0.117	0.124	0.059	0.071	0.045	0.127	0.184	-0.046	-0.042	-0.186	0.047
N-gram stats																
NovelNgram_4	0.358	0.386	0.516	0.600	0.277	0.280	0.295	0.283	-0.231	-0.221	0.282	0.285	0.018	0.088	0.073	0.107
NovelNgram_3	0.355	0.390	0.494	0.591	0.290	0.276	0.300	0.291	-0.235	-0.219	0.286	0.289	0.071	0.105	0.107	0.118
NovelNgram_2	0.337	0.384	0.439	0.570	0.276	0.252	0.298	0.292	-0.208	-0.191	0.283	0.287	0.064	0.093	0.170	0.156
NovelNgram_1	0.286	0.349	0.282	0.410	0.123	0.114	0.271	0.267	-0.070	-0.087	0.229	0.242	0.111	0.119	0.158	0.178
Length (no. tokens)	0.247	0.207	0.263	0.277	0.096	0.099	0.048	0.044	-0.008	0.004	0.230	0.208	-0.133	-0.144	-0.171	-0.184
Contextual Embedding																
BERTScore P	0.221	0.237	-0.095	-0.051	0.049	0.052	0.336	0.320	0.152	0.125	0.245	0.266	**0.186**	**0.208**	0.022	0.030
BERTScore R	0.327	0.360	0.026	0.015	0.171	0.158	0.335	0.340	0.139	0.126	0.426	0.415	0.131	0.135	0.078	0.095
BERTScore F1	0.304	0.329	-0.041	-0.020	0.107	0.100	0.378	0.375	0.167	0.144	0.360	0.367	0.174	0.186	0.049	0.072
BERTScore P_art	0.465	0.513	0.493	0.548	0.350	0.338	0.449	0.429	0.351	0.300	0.443	0.422	0.176	0.196	-0.028	-0.026
BERTScore R_art	0.395	0.426	0.452	0.497	0.175	0.180	0.230	0.215	0.180	0.145	0.344	0.326	0.046	0.069	-0.049	-0.053
BERTScore F1_art	0.464	0.514	0.493	0.556	0.295	0.292	0.381	0.358	0.299	0.246	**0.447**	**0.423**	0.137	0.157	-0.054	-0.048
Neural entailment																
FactCC	0.438	0.492											0.072	0.072		
Dep Entail	0.447	0.440											0.113	0.058		
Q&A based																
FEQA	-0.010	-0.018											0.008	0.026		
QAGS	0.267	0.314	0.382	0.466									-0.007	-0.022	**0.203**	**0.217**
QAEval-F1 (Deutsch et al., 2021a)											.300	.290				
Text Generation based																
CBMI (BART_base + cnn)*	0.076	0.099	0.040	0.133	0.222	0.194	-0.013	-0.045	0.082	0.030	0.103	0.069	-0.095	-0.113	-0.058	-0.022
BARTScore (BART-large + cnn)**	0.530	0.561	0.613	0.673	0.262	0.249	0.459	0.429	0.278	0.231	0.390	0.363	0.168	0.174	0.097	0.080
BARTScore (BART-large + cnn + para)**	0.507	0.543	0.548	0.624	0.343	0.328	0.438	0.419	0.350	0.305	0.422	0.385	0.177	0.175	0.041	0.046
Proposed																
HaRiM (BART_large + cnn)	**0.542**	**0.581**	**0.620**	**0.679**	0.336	0.317	**0.463**	**0.437**	0.321	0.268	0.414	0.391	0.167	0.175	0.101	0.087
HaRiM (BART-large + cnn + para)	0.515	0.556	0.530	0.610	**0.387**	**0.356**	0.423	0.408	**0.366**	**0.314**	0.426	0.390	0.173	0.172	0.037	0.042

Table B.3: Metric-Human correlation (segment-level) in Spearman's r and Pearson's ρ. The best performance are bolded and second-bests are underlined.

score	r	ρ
$\log(\mathrm{H}_{lm}/\mathrm{H}_{\mathrm{s2s}})$	0.05	0.05
$\log(\mathrm{H}_{lm}/\mathrm{H}_{\mathrm{s2s}})\_len$	0.05	0.05
$\mathrm{H}_{lm}/\mathrm{H}_{\mathrm{s2s}}$	0.05	0.05
$(\mathrm{H}_{lm}/\mathrm{H}_{\mathrm{s2s}})\_len$	0.05	0.05
$\mathrm{H}_{s2s} * \mathrm{H}_{\mathrm{lm}}$	0.23	0.10
$(\mathrm{H}_{s2s} * \mathrm{H}_{\mathrm{lm}})\_len$	0.00	-0.01
$\log(\mathrm{H}_{s2s} * \mathrm{H}_{\mathrm{lm}})\_len$	0.00	0.01
$(\mathrm{H}_{lm} - \mathrm{H}_{\mathrm{s2s}})\_len$	0.04	0.04
H_{lm}	0.22	0.17
H_{lm}_len	0.04	0.02
H_{s2s}	0.22	0.19
H_{s2s}_len	-0.03	-0.02
-HaRiM_lmless	**0.46**	**0.50**
-HaRiM	**0.46**	**0.50**
-HaRiM (quintic) _lmless	0.45	0.40
-HaRiM (quintic)	0.45	0.40
-HaRiM_top5mean	0.04	0.06
-HaRiM_bot5mean	0.14	0.17

Table B.4: Variation tested over FRANK *CNNDailyMail* split. H denotes entropy. *_len* refers to length normalization. Entropy-based scores are performing worse. We also tested other variations for aggregating token-level scores into a scalar such as idf, tf-idf reweighting of HaRiM (not presented here) which do nothing more than worsening the correlation to human judgements similarly to top/bot 5 averaging.

Kendall's τ	Abstractive 1200 outputs				Extractive 400 outputs			
Metrics	**Con**	**Coh**	**Flu**	**Rel**	**Con**	**Coh**	**Flu**	**Rel**
N-gram matching								
ROUGE 1	0.117	0.129	0.057	0.219	0.094	0.209	0.063	0.161
ROUGE 2	0.107	0.128	0.041	0.173	0.066	0.153	0.030	0.118
ROUGE L	0.114	0.096	0.071	0.180	0.063	0.164	0.033	0.123
METEOR	0.094	0.091	0.025	0.217	0.003	0.148	0.121	0.201
sacreBLEU	0.109	0.201	0.103	0.234	0.022	0.070	0.091	0.147
ROUGE 1_art	0.050	-0.021	-0.005	0.114	0.104	0.117	0.109	0.066
ROUGE 2_art	0.150	0.020	0.073	0.144	0.112	0.129	0.113	0.077
ROUGE L_art	0.157	0.045	0.083	0.157	0.123	0.166	0.088	0.089
METEOR_art	0.066	-0.043	0.024	0.082	0.107	0.087	0.096	0.033
sacreBLEU_art	0.023	-0.016	-0.036	0.115	0.098	0.123	0.101	0.078
N-gram stats								
NovelNgram_4	0.241	0.230	0.245	0.214	0.042	0.085	0.140	0.166
NovelNgram_3	0.305	0.238	0.250	0.217	0.042	0.085	0.147	0.170
NovelNgram_2	**0.315**	0.243	0.223	0.218	0.045	0.084	0.140	0.168
NovelNgram_1	0.299	0.229	0.088	0.189	0.040	0.088	0.082	0.154
Length (no. tokens)	-0.015	-0.039	-0.097	0.120	0.050	0.150	0.068	0.137
Contextual Embedding								
BERTScore P	0.092	0.316	0.135	0.229	0.043	0.019	0.124	0.166
BERTScore R	0.124	0.257	0.071	**0.309**	0.020	0.168	0.154	**0.239**
BERTScore F1	0.110	0.330	0.124	0.288	0.040	0.085	0.154	0.229
BERTScore P_art	0.263	0.334	0.225	0.317	0.110	0.189	0.187	0.234
BERTScore R_art	0.102	0.139	0.070	0.239	0.083	0.141	0.141	0.160
BERTScore F1_art	0.208	0.266	0.164	0.319	0.112	**0.196**	0.184	0.225
Text Generation based								
CBMI (BART_base + cnn)*	0.089	-0.114	-0.030	0.016	0.066	0.099	-0.068	0.028
BARTScore (BART_large + cnn)**	0.222	**0.368**	0.188	0.288	0.099	0.102	**0.191**	0.178
BARTScore (BART_large + cnn + para)**	0.281	0.350	0.249	0.303	0.128	0.111	0.188	0.180
Proposed								
HaRiM^+ **(BART_large + cnn)**	0.278	0.366	0.219	0.308	0.098	0.120	0.185	0.190
HaRiM^+ (BART_large + cnn + para)	0.306	0.339	**0.260**	0.306	**0.126**	0.110	0.176	0.183

Table B.5: Metric-to-human judgement correlation (segment-level) reported in Kendall's τ. **Bold**-face values are the largest correlating metrics, underlined are second-large values amongst the metrics. Hallucination Risk(HaRiM^+) outperforms others in most criteria. We provide permutation test result in Appendix. *(Wu et al., 2021), **(Yuan et al., 2021)

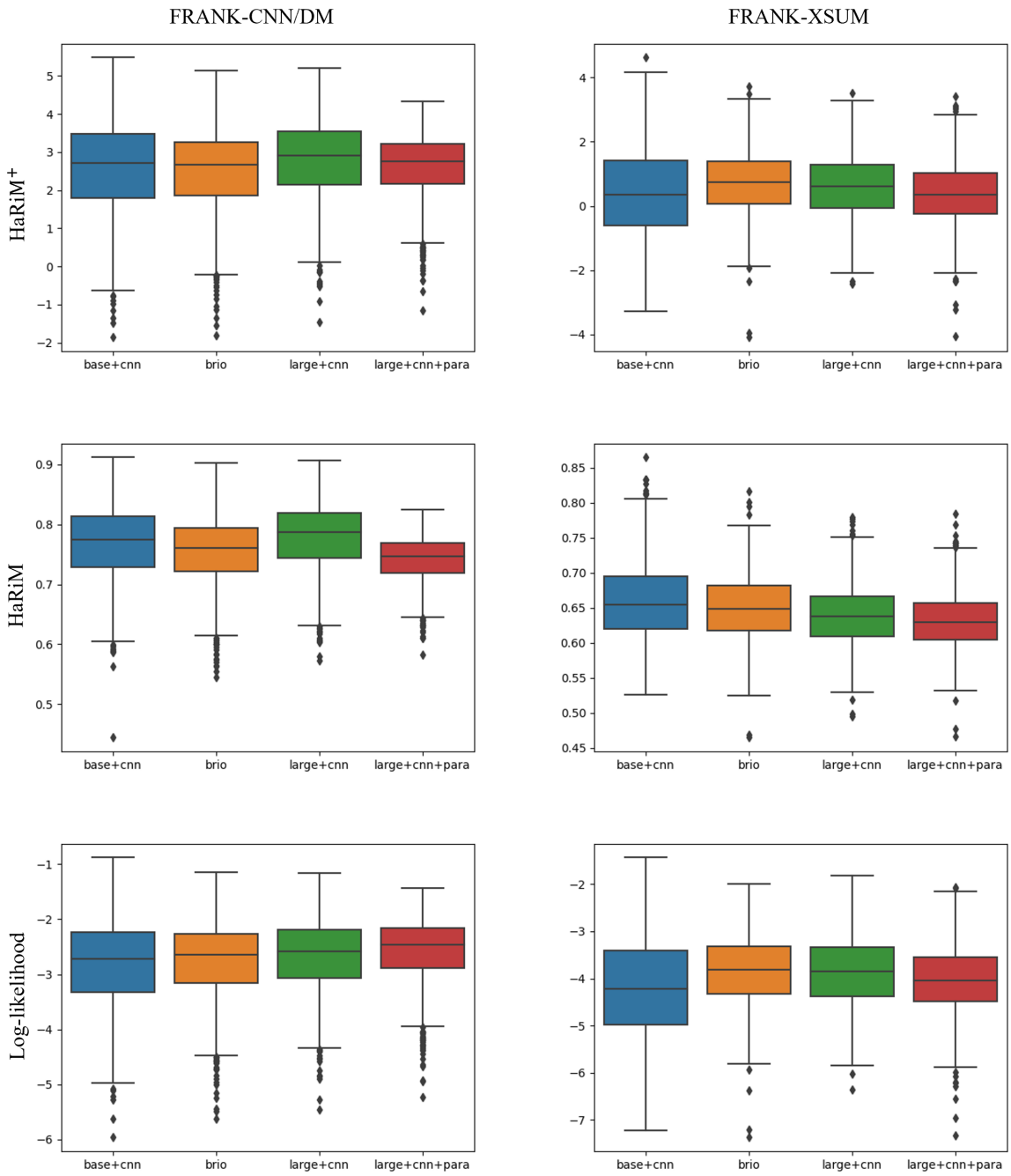

Figure A.5: Boxplot of HaRiM and log-likelihood scales, varying with the evaluating summarizer weight. `base+cnn`: BART-base fine-tuned on *CNN/DailyMail*, `brio`: BRIO (Meng et al., 2021), `large+cnn`: BART-large fine-tuned on *CNN/DailyMail*, `large+cnn+para`: further fine-tuned checkpoint of the previous model on ParaBank2 corpus as suggested in (Yuan et al., 2021).

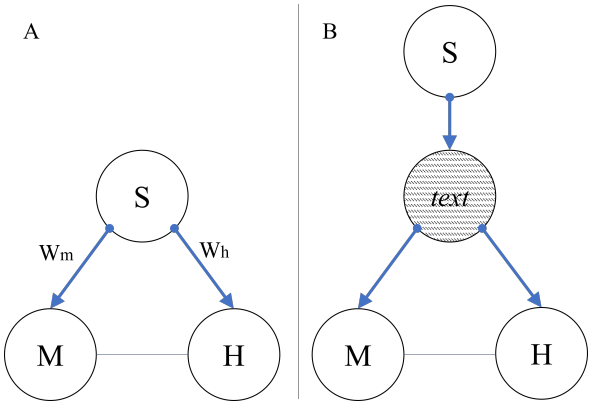

Figure A.6: Graphical model representation attributing to the factors that affects metric (M)-human (H) correlation. A is the graphical model that supports the use of partial correlation as argued in (Pagnoni et al., 2021). B is the graphical model that adheres to our argument that why should we measure correlation, ignoring the effect of the generation system (S) whose effect is hindered by observed child node, $text$.

Source Article			
A youngster has emulated Barcelona star Martin Montoya and scored an audacious 27-yard goal into a basketball hoop - twice. Schoolboy Frankie Franz watched the Spanish right-back pull off the staggering trick shot in a video recorded at Barcelona's Ciutat Esportiva training ground earlier in the month. The viral clip shows the 23-year-old defender lifting the ball into the net to the sound of gasps from his team mates at the Catalonia club. Joking that he could do the same with his mum and grandmother, nine-year-old Frankie, who is an academy player with Dagenham and Redbridge Football Club, took to the garden to have a go. He moved the basketball hoop into the middle of the goal and after a little run up sent the ball straight through the net first time. In the video he can be seen turning to face the camera looking absolutely gob-smacked as he places his hands behind his head. A slow motion look at the clip captures the trajectory of the ball, which bounces off the backboard and goes through the hoop below. His mother Lucy, 32, of Upminster, east London, said: 'He loves football and after he saw the Barcelona player do the show he said 'I'll be able to do that'. (...) The talented young centre midfielder has played football since he could walk and dreams of one day turning out for Real Madrid or Barcelona. The young centre midfielder is an academy player with Dagenham and Redbridge Football Club and dreams of one day turning out for Real Madrid or Barcelona.			
Model	**Summary**	HaRiM$^+$ **Score** ↑	**Score Gain** ↑
Reference	frankie franz watched the right-back pull off the audacious shot in a video . nine-year-old joked with his mum and grandmother that he could make it . youngster moved hoop into middle of the garden and twice achieved feat . frankie is an academy player with dagenham and redbridge football club . he plays centre midfield and dreams of one day turning out for barcelona .	2.5723	-
Self-generation (BART-large+cnn)	frankie franz watched the spanish right-back pull off the staggering trick shot in a video recorded at barcelona's ciutat esportiva training ground earlier in the month. the viral clip shows the 23-year-old defender lifting the ball into the net to the sound of gasps from his team mates at the catalonia club. joking that he could do the same with his mum and grandmother, frankie took to the garden to have a go. he moved the basketball hoop into the middle of the goal and after a little run up sent the ball straight through the net first time.	4.5318	+1.9595
BottomUpSummary (Factuality=0.0)	frankie franz watched the spanish right-back pull off the trick shot in a video recorded at barcelona 's catalonia club . the 23-year-old defender took to the garden to have a go and moved the basketball hoop into the net to the goal . his mother lucy , 32 , said : ' me said ' i will be able to do . ' .	1.3673	-1.2050
Reference (w/ wrong subject)	martin montoya watched the right-back pull off the audacious shot in a video . nine-year-old joked with his mum and grandmother that he could make it . youngster moved hoop into middle of the garden and twice achieved feat . frankie is an academy player with dagenham and redbridge football club . he plays centre midfield and dreams of one day turning out for barcelona .	2.5595	-0.0128
Reference (w/ negation)	frankie franz did not watch the right-back pull off the audacious shot in a video . nine-year-old joked with his mum and grandmother that he could make it . youngster moved hoop into middle of the garden and twice achieved feat . frankie is an academy player with dagenham and redbridge football club . he plays centre midfield and dreams of one day turning out for barcelona .	2.3178	-0.2545

Table B.6: Testing HaRiM$^+$ metric under hallucination detecting scenario. Part of the source article, which is irrelevant to the summaries are omitted for clarity. The words highlighted red are hallucinated information deliberately injected to the reference. BottomUpSummary refers to abstractive summarization system suggested in (Gehrmann et al., 2018).

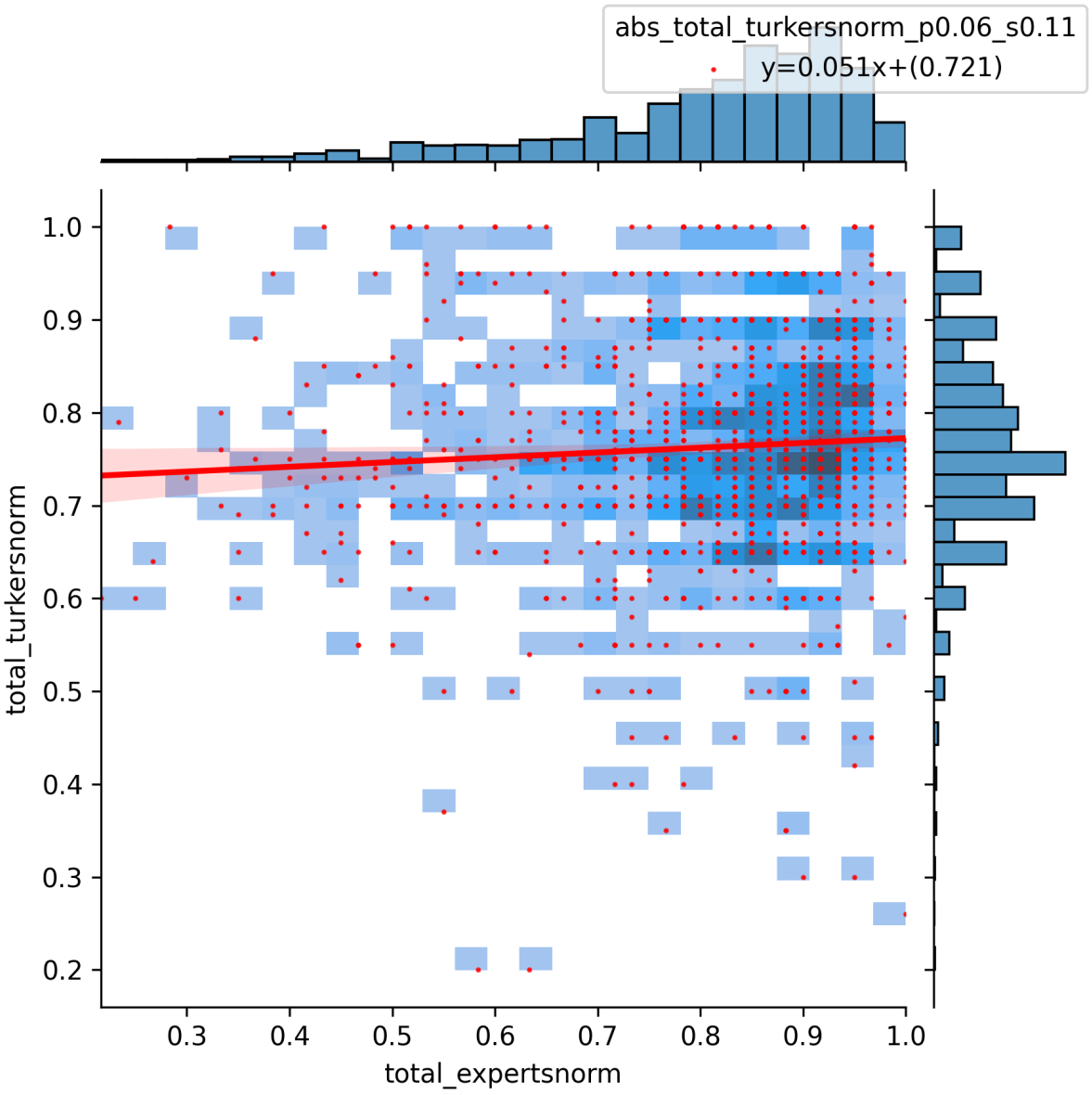

Figure A.7: Averaged experts' judgements vs. Averged turkers' judgements on SummEval, (datapoints are outputs from **abstractive** summarization models)

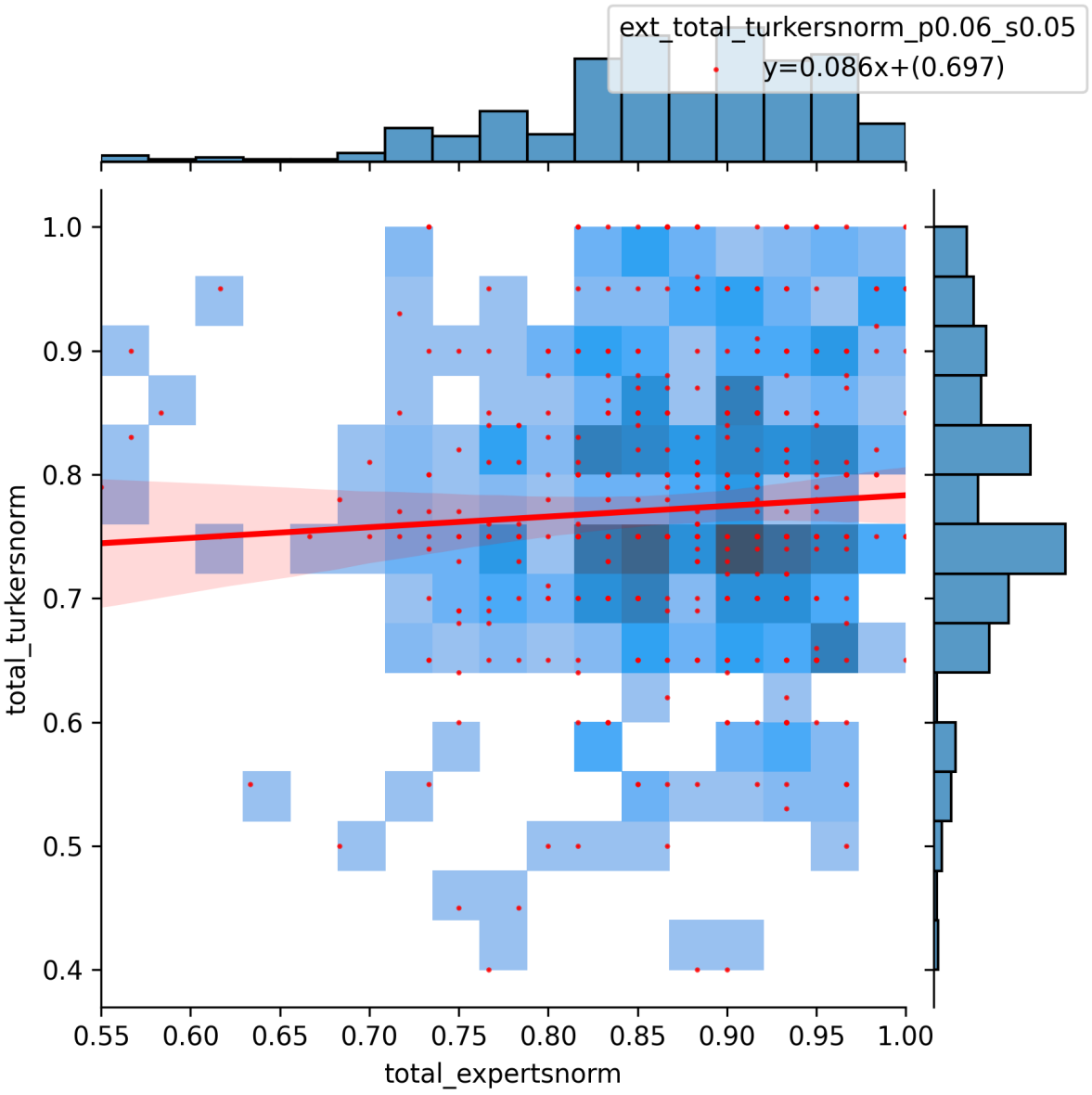

Figure A.8: Averaged experts' judgements vs. Averged turkers' judgements on SummEval, (datapoints are outputs from **extractive** summarization models)

Source Article			
The view that Manchester City's chance at defending their Premier League title has been ruined through bad spending gathered pace after they were defeated by a club whose entire team cost less than half one of their substitutes. Crystal Palace's XI on Monday night may only have been worth a mere £17m, but left back Martin Kelly still made it through a City defence deemed good enough to keep £40m signing Eliaquim Mangala on the bench to tee up a chance for Wilfried Zaha just 60 seconds into the game. Mangala joined from Porto in August last year and is contracted to City until June 2019. Eliaquim Mangala (green bib) prepares to come on but he never made it off the Manchester City bench However, striker Glenn Murray succeeded in putting another dent in City's chances of redeeming themselves after a run of four losses away, when he scored Palace's first goal. Murray cost Palace nothing when joined from arch rivals Brighton in 2011. Jason Puncheon, signed for a comparative pittance of £1.9m, delivered City their final blow with a goal from a finely executed free-kick. Glenn Murray (left) cost Palace nothing four years ago yet found a way past the City defence Another expensive City player, £24m-man Yaya Toure, got his team back in the game with 12 minutes left, but they couldn't penetrate Palace's defence to find an equaliser and a 2-1 defeat leaves them nine points adrift of the top. Toure joined from Barcelona in July 2010 and is contracted to City until 2017. After spending a total of £500m pounds on transfer fees, City might have expected to be higher than a precarious fourth in the league, but judging by their latest results, it's teams like Crystal Palace that seem to be getting their value for money. Mangala has endured a miserable first season at the Etihad Stadium since his £40million move			
Model	**Summary**	HaRiM$^+$ **Score** ↑	**Score Gain** ↑
Reference	manchester city beaten 2-1 by crystal palace on easter monday . 40m signing eliaquim mangala was left on the bench . crystal palace 's entire starting xi cost just 17million . click here for all the latest manchester city news .	0.8913	-
Self-generation (BART-large+cnn)	manchester city lost 2-1 to crystal palace at the etihad on monday night. crystal palace's entire team cost less than half one of manchester city's substitutes. eliaquim mangala and yaya toure were both left on the bench. city have spent a total of £500m on transfer fees so far this season.	3.7006	+2.8093
BottomUpSummary (Factuality=0.0)	crystal palace 's xi is contracted to city until june 2019 . jason puncheon signed for 1.9 m from porto in august last year . glenn murray has scored four goals in the premier league .	-0.4833	-1.3746
Reference (w/ wrong subject)	manchester city beaten 2-1 by crystal palace on easter monday . 40m signing wilfried zaha was left on the bench . crystal palace 's entire starting xi cost just 17million . click here for all the latest manchester city news .	0.5746	-0.3167
Reference (w/ negation)	manchester city beaten 2-1 by crystal palace on easter monday . 40m signing eliaquim mangala was not left on the bench . crystal palace 's entire starting xi cost just 17million . click here for all the latest manchester city news .	0.7715	-0.1198

Table B.7: Testing HaRiM$^+$ metric under hallucination detecting scenario. Part of the source article irrelevant to the summaries are omitted for clarity. The words highlighted red are hallucinated information deliberately injected to the reference. BottomUpSummary refers to abstractive summarization system suggested in (Gehrmann et al., 2018).

Source Article
(CNN)Soon, America will be too fat to fight. Forget about rampant diabetes, heart attacks and joint problems – the scariest consequence arising out of our losing battle with the bulge is the safety of our country. In about five years, so many young Americans will be grossly overweight that the military will be unable to recruit enough qualified soldiers. That alarming forecast comes from Maj, Gen. Allen Batschelet, who is in charge of U.S. Army Recruiting Command. Obesity, he told me, “is becoming a national security issue." I was so taken aback by Batschelet’s statement that I felt the need to press him. Come on! Obesity? A national security crisis? The General didn’t blink. “In my view, yes." Of the 195,000 young men and women who signed up to fight for our country, only 72,000 qualified. Some didn’t make the cut because they had a criminal background, or a lack of education, or too many tattoos. But a full 10% didn’t qualify because they were overweight. Before you accuse me of sensationalizing, it’s that 10% figure that worries General Batschelet the most. “The obesity issue is the most troubling because the trend is going in the wrong direction," he said. “We think by 2020 it could be as high as 50%, which mean only 2 in 10 would qualify to join the Army." He paused. “It’s a sad testament to who we are as a society right now." The problem is so worrisome for the Army that recruiters have become fitness coaches, like the trainers on the NBC show, “The Biggest Loser." Yes, your tax dollars pay for Army recruiters to play Dolvett Quince or Jillian Michaels to whip could-be recruits into shape with the hope they can diet and exercise their way to become real recruits. If they lose enough weight, they’re sent to boot camp. Some make it; many don’t. But, General Batschelet told me the Army must try. “We are the premier leader on personal development in the world," he told me. “We want to see you grow and become a leader. That is a great strength in our Army." Except the Army never considered the type of growth it’s now contending with. Nowadays “personal development" means working on both character and ... girth. The general, along with so many others in this country, is struggling with why so many Americans, despite all the warnings, continue to eat too much and exercise too little. I have a theory. It ain’t pretty. But it’s got to be true: We just don’t care. “The acceptance of obesity is prevalent," according to Claire Putnam, an obstetrician and gynecologist who believes obesity is a national crisis right now. “When you look around you, 70% of adults are overweight or obese. It’s seems normal," she said. Just look at the numbers: More than one-third of U.S. adults are obese. Seventeen percent of all children and adolescents in the U.S. are obese. That’s triple the rate from just a generation ago. So, maybe we should face the fact that we’ve grown comfortable with our girth. It is crystal clear we haven’t the foggiest idea of who needs to lose weight and who doesn’t. Just the other day, Twitter trolls scolded the singer, Pink, for gaining weight. Pink is not remotely fat. Neither is Selena Gomez, haters. Or Britney Spears, hecklers. If 70% of us are overweight in this country, why are there so many willing to fat-shame people who are not remotely obese? Maybe it’s easier to criticize others for carrying extra weight than to admit we have a weight problem ourselves. Because it is abundantly clear we are wallowing in denial. Dr. Putnam points to one of Kaiser Permanante’s medical questionnaires. You know, the paperwork patients are asked to fill out before they see the doctor. There is actually a box on the form that allows the patient to “opt out of talking about obesity." Some patients refuse to step on the scale. “You want to be sensitive to that patient," Putnam told me. “You don’t want to nag. But, doctors need to step in and say we need to fix this." CNN’s chief medical correspondent, Dr. Sanjay Gupta, agrees with Putnam. “Perceptions of weight are a big part of the problem," he said to me. “If a person is overweight – as difficult as it is – they ought to be told. You know, this issue reminds me of the issue with concussions. We should call them what they really are: a brain injury, not ’getting your bell rung.’ In the same vein, we should tell people who are overweight or obese that, clinically, they’re ’overweight’ or ’obese’ and at risk for just about every chronic disease in the book." In other words, chubby is not the proper way to describe a person who is obese. Just like “fat" is not the proper term for Pink or Selena Gomez. And, yes, semantics matter. According to the CDC, 81% of overweight boys and 71% of overweight girls believe they are just the right weight. We’ve clearly lost our perspective on what’s normal when it comes to a healthy weight. So much so it’s becoming a national security problem. So what will it take? The answer cannot be the U.S Army.

Model	Summary	HaRiM$^+$ Score ↑	Score Gain ↑
Reference	in a few years , the military will be unable to recruit enough qualified soldiers because of america ’s obesity problem . carol costello : we have a serious national security issue at hand , but it ’s within our control if we could own up to it .	1.0219	-
Self-generation (BART-large+cnn)	of the 195,000 young men and women who signed up to fight for our country, only 72,000 qualified. a full 10% didn’t qualify because they were overweight. “it’s a sad testament to who we are as a society right now," says maj, gen. allen batschelet.	5.2130	+4.1911
BottomUpSummary (Factuality=0.0)	many young americans will be overweight that the military will be able to recruit enough soldiers . gen. allen batschelet is a national security issue for the u.s. army . he says the obesity issue is so many that it ’s too fat to fight .	0.7128	-0.3091
Reference (w/ wrong subject)	in a few years , the military will be unable to recruit enough qualified soldiers because of america ’s obesity problem . claire putnam : we have a serious national security issue at hand , but it ’s within our control if we could own up to it .	1.0111	-0.0108
Reference (w/ negation)	in a few years , the military will be unable to recruit enough qualified soldiers because of america ’s obesity problem . carol costello : we do not have a serious national security issue at hand , but it ’s within our control if we could own up to it .	0.9572	-0.0647

Table B.8: Testing HaRiM$^+$ metric under hallucination detecting scenario. Part of the source article irrelevant to the summaries are omitted for clarity. The words highlighted red are hallucinated information deliberately injected to the reference. BottomUpSummary refers to abstractive summarization system suggested in (Gehrmann et al., 2018).

Source Article
It's well known that exercise can make your muscles bigger. Now, a study has found it may make your brain larger, too. Physical activity can increase grey matter in the brain, increasing the size of areas that contribute to balance and coordination, according to Health Day news. The changes in the brain may have health implications in the long-term, such as reducing the risk of falling, said the study's author, Dr Urho Kujala, of the University of Jyvaskyla. Scroll down for video Exercise can increase the size of areas of the brain that contribute to balance and coordination, a study found It could also reduce the risk of being immobile in older age, he added. Dr Kujala said physical activity has already been linked to a number of health benefits, such as lower levels of body fat, reduced heart disease risk factors, better memory and thinking, and a lower risk of type 2 diabetes. But he and his team wanted to understand how exercise affects the brain. They recruited 10 pairs of identical twins, who were all men aged 32 to 36 years. Focusing on twins, who have the same DNA, would allow researchers to see how their environment affects their bodies. In each pair of twins, one brother had exercised more over the past three years than the other, though they reported they carried out similar levels of exercise earlier in their lives. Dr Kujala said: 'On average, the more active members of twin pairs were jogging three hours more per week compared to their inactive co-twins.' The twins had MRI scans of their brains so researchers could see whether physical activity had any impact on the size of their brains, and specific regions. Exercise didn't seem to affect the size of the brain as a whole, Dr Kujala said. But there was a connection between more activity and more brain volume in areas related to movement, he added. Previous research found exercise is linked to lower levels of body fat, a reduced risk of heart disease, better memory and thinking, and a lower risk of type 2 diabetes The twins who exercised more did a better job of controlling their blood sugar, which reduces the risk of diabetes, a finding which is already well-known. The study was published in the journal Medicine & Science in Sports & Exercise. It comes after US researchers found regular exercise can also make you smarter. University of South Carolina experts found regular treadmill sessions create more mitochondria - structures in the cells that produce the body's energy - in the brain. This energy boost helped the brain to work faster and more efficiently, effectively keeping it younger, researchers said. In the short term this could reduce mental fatigue and sharpen your thinking in between gym sessions. And building up a large reservoir of mitochondria in the brain could also create a 'buffer' against age-related brain diseases such as Alzheimer's.

Model	Summary	HaRiM$^+$ Score ↑	Score Gain ↑
Reference	study : exercising increases the amount of grey matter in the brain . it makes areas of the brain that control balance and co-ordination bigger . in the long term this could reduce the risk of falling or becoming immobile . previous studies show exercise can stave off alzheimer 's and diabetes .	2.1515	-
Self-generation (BART-large+cnn)	physical activity can increase grey matter in the brain, a study found. it can increase the size of areas that contribute to balance and coordination. changes may have health implications in the long-term, such as reducing the risk of falling, said the study's author, dr urho kujala, of the university of jyvaskyla.	5.1145	+2.9630
BERTSum (Factuality=1.0)	exercise can increase grey matter in the brain , increasing the size of areas that contribute to balance and coordination . study 's author , dr urho kujala , of the university of jyvaskyla , said physical activity has already been linked to a number of health benefits , such as lower levels of body fat , reduced heart disease risk factors , better memory and thinking , and a lower risk of type 2 diabetes .	3.8029	+1.6514
Reference (w/ wrong subject)	study : exercising increases the amount of mitochondria in the brain . it makes areas of the brain that control balance and co-ordination bigger . in the long term this could reduce the risk of falling or becoming immobile . previous studies show exercise can stave off alzheimer 's and diabetes .	1.9037	-0.2478
Reference (w/ negation)	study : exercising does not increase the amount of grey matter in the brain . it makes areas of the brain that control balance and co-ordination bigger . in the long term this could reduce the risk of falling or becoming immobile . previous studies show exercise can stave off alzheimer 's and diabetes .	1.9733	-0.1782

Table B.9: Testing HaRiM$^+$ metric under hallucination detecting scenario. Part of the source article irrelevant to the summaries are omitted for clarity. The words highlighted red are hallucinated information deliberately injected to the reference. BERTSum refers to extractive summarization system suggested in (Liu and Lapata, 2019).

Source Article
The respected law professor from Philadelphia now being investigated after allegedly emailing students a link to pornographic footage, was once a contestant on Who Wants to Be a Millionaire, it has emerged. Lisa McElroy, a 50-year-old Drexel professor, appeared on the show in 2010 while it was still hosted my Meredith Vieira. And like her apparent March 31 email mishap, her game show appearance ended with a very public mistake. McElroy, who teaches legal writing, got tripped up on the $12,500 level after flying through the first few questions, notes Philly.com. Wishes she was a millionaire: Drexel law profesor professor Lisa McElroy allegedly sent a link to a pornographic website to her students. In 2010, she appeared on the TV game show Who Wants to Be a Milionaire Mother of two: The mother of two shared an anecdote with then-host Meredith Vieira about having to scramble to find a babysitter for her kids and someone to teach her class after learning she was to appear on the show just two days before taping Lost it: McElroy was tripped up on the $12,500 question. Despite having used two lifelines, she answered wrong and walked away with around $5,000 The questions read: 'As a result of General Motor's bankruptcy declaration in 2009, what foreign government became one of its largest shareholders?' Even after using two of her lifelines to narrow down the answer, McElroy answered China, which was incorrect. The correct answer was Canada. She walked away with around $5,000. McElroy, who is a children's book and biography author, is apparently also a mother. She opened the appearance by sharing an anecdote with Vieira about having to scramble to find a babysitter after being informed she was chosen to be on Millionaire jsut two days prior to taping. She's accused of sending the inappropriate message this past March 31 under the subject line: 'Great article on writing briefs.' However, when recipients opened the enclosed link, philly.com reports that they were directed to a video of 'a woman engaging in a sexually explicit act'. Lisa McElroy, 50, who teaches legal writing at Drexel University, reportedly sent the inappropriate message on March 31 baring the subject line: 'Great article on writing briefs' Following a number of complaints, the college issued an apology to students. The message read: 'As you may be aware, some students erroneously received an email this morning directing them to a... post that included some inappropriate material. 'We take this matter seriously and apologize for any upset it may have caused.' The university says federal law requires it investigate all reports of inappropriate behaviors of a sexual nature. McElroy did not immediately respond to an email sent to her university account by the Associated Press. When recipients opened the enclosed link, philly.com reports that they were directed to a video of 'a woman engaging in a sexually explicit act' It's not the first time the married mother-of-two has appeared in the spotlight. She is also an accomplished author with a number of published biographies and children's books. On her website, www.lisamcelroy.com, she describes herself as a 'Supreme Court junkie.' She adds that her favorites ways of relaxing include 'crawling under the covers with a dog or two and a really good book' or 'hanging out' with her two adolescent daughters. Regarding the recent email scandal, David Lat - a lawyer and legal commenter -suggests she could have been 'hacked' or made a 'copy/paste error'. While an internal investigation gets underway, it's been reported that McElroy has been placed on administrative leave. While an internal investigation gets underway, it's been reported that McElroy has been placed on administrative leave from Drexel University (seen above)

Model	Summary	HaRiM^{+} Score ↑	Score Gain ↑
Reference	lisa mcelroy , 50 , who teaches legal writing at drexel university , reportedly sent the ' inappropriate ' message on march 31 . when recipients clicked the enclosed link , they were allegedly directed to a video of ' a woman engaging in a sexually explicit act ' . mcelroy appeared on the popular game show in 2010 with then-host meredith vieira but lost the game after reaching just $ 12,500 . along with teaching law , mcelroy is also an accomplished author with a number of published biographies and children 's books . has been placed on leave while school investigates .	2.3270	-
Self-generation (BART-large+cnn)	lisa mcelroy, a 50-year-old drexel professor, appeared on the show in 2010 while it was still hosted my meredith vieira. she's accused of sending the inappropriate message this past march 31 under the subject line: 'great article on writing briefs' when recipients opened the enclosed link, philly.com reports that they were directed to a video of 'a woman engaging in a sexually explicit act' the married mother-of-two has been placed on administrative leave.	4.9714	+2.6444
BERTSum (Factuality=1.0)	lisa mcelroy , 50 , who teaches legal writing at drexel university , appeared on the show in 2010 while it was still hosted my meredith vieira . she got tripped up on the $ 12,500 level after flying through the first few questions , philly.com reports . mcelroy answered wrong and walked away with around $ 5,000 .	3.2028	+0.8758
Reference (w/ wrong subject)	lisa mcelroy , 50 , who teaches legal writing at philadelphia university , reportedly sent the ' inappropriate ' message on march 31 . when recipients clicked the enclosed link , they were allegedly directed to a video of ' a woman engaging in a sexually explicit act ' . mcelroy appeared on the popular game show in 2010 with then-host meredith vieira but lost the game after reaching just $ 12,500 . along with teaching law , mcelroy is also an accomplished author with a number of published biographies and children 's books . has been placed on leave while school investigates .	2.2122	-0.1148
Reference (w/ negation)	lisa mcelroy , 50 , who teaches legal writing at drexel university , reportedly did not send the ' inappropriate ' message on march 31 . when recipients clicked the enclosed link , they were allegedly directed to a video of ' a woman engaging in a sexually explicit act ' . mcelroy appeared on the popular game show in 2010 with then-host meredith vieira but lost the game after reaching just $ 12,500 . along with teaching law , mcelroy is also an accomplished author with a number of published biographies and children 's books . has been placed on leave while school investigates .	2.2022	-0.1248

Table B.10: Testing HaRiM^{+} metric under hallucination detecting scenario. Part of the source article irrelevant to the summaries are omitted for clarity. The words highlighted red are hallucinated information deliberately injected to the reference. BERTSum refers to extractive summarization system suggested in (Liu and Lapata, 2019).