Asia Pacific Science Press

Journal of Advances in Engineering and Technology
ISSN(O):3005-7507
ISSN(P):3005-7493
Vol. 3 No.4 (2026)

When More Metrics Do Not Improve Faithfulness Assessment: A Document-Level Evaluation of AI-Generated News Summaries

Cheng Chen[1], Alina Andreeva[2], Ramin Jalalpoor[3], Jinjiang Liu[4], Chaofan Yue[5], Zhiwei Cheng[2], Cuong Huynh[5], Zyad Abdullah Al-Shuja[5], Xiangyu Li[6]*

1. School of Media and Art, Nanchang Institute of Technology, Nanchang, Jiangxi, 330099, China
2. We Young Open Lab, Hangzhou We Young Geron Technology Co., Ltd., Hangzhou, Zhejiang, 310000, China
3. Faculty of Robotics and Artificial Intelligence, ITMO University, St. Petersburg, 197101, Russia
4. Faculty of Physics, St. Petersburg State University, St. Petersburg, 198504, Russia
5. Faculty of Control Systems and Robotics, ITMO University, St. Petersburg, 197101, Russia
6. Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China

Corresponding author: Xiangyu Li, xiangyuli@sjtu.edu.cn

Copyright: 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY-NC 4.0), permitting distribution and reproduction in any medium, provided the original author and source are credited, and explicitly prohibiting its use for commercial purposes.

Abstract: Automatic scores are attractive for screening artificial intelligence (AI)-generated news summaries, but their agreement with human judgments must be evaluated without sharing source articles between model fitting and testing. We examined whether a fixed linear fusion of five published scores improves the ranking of human-rated source faithfulness over the published entailment score alone. We reanalysed the Creative Commons licensed XSum hallucination-annotation release, using its existing human judgments and precomputed ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, and entailment scores. Before inspecting outcomes, we fixed mechanical eligibility rules, a deterministic document-level 80/20 split, ridge regression with penalty 1.0, Spearman correlation as the primary endpoint, and 2,000 document-cluster bootstrap draws. The eligible file contained 1,992 system-summary records from 498 BBC article IDs; 399 articles and 1,596 records were used for fitting, and 99 articles and 396 records for testing. On the held-out set, fusion correlated with human faithfulness at $\rho = 0.425$ (95% bootstrap interval 0.325 to 0.515), versus $\rho = 0.445$ (0.342 to 0.532) for entailment. The prespecified difference was 0.020 (0.100 to 0.066). For the secondary factuality-vote target, correlations were 0.285 and 0.285 after rounding, with no detectable fusion advantage. This reanalysis shows no evidence that this fixed fusion improves ranking over entailment in the released sample. The labels do not assess source attribution or editorial fidelity, and no new human annotation or summary generation was performed.

Keywords: News Summarization; Factual Consistency; Source Faithfulness; Human Annotation; Metric Fusion; Document-Level Validation

Published: Nov 12, 2026

DOI: https://doi.org/10.62177/jaet.v3i4.1811

1. Introduction

Artificial intelligence (AI)-generated summaries can make a news article easier to scan, but the compression may introduce

claims unsupported by the article. In a journalism workflow, a summary that reads fluently is not necessarily a reliable account of the source. XSum formalised an extreme summarization task using BBC news articles and short, highly abstractive target summaries [1]. Subsequent human evaluation distinguished source faithfulness from factuality and reported substantial hallucination in generated XSum summaries [2]. These findings motivate an engineering question: how well can an automatic score order summaries according to existing human source-based judgments?

The measurement target matters. Source faithfulness concerns support in the supplied article. A statement can be plausible or even true elsewhere while remaining unsupported by that article. Factuality in the released annotation set is a separate human vote about whether a hallucinated summary is factual [2]. Source attribution would require a label about who or what is credited for each claim; editorial fidelity would require criteria for context, emphasis, and framing. Neither of those latter labels is supplied in the selected score file. Accordingly, the present study tests source faithfulness as its primary target and treats the released factuality vote as secondary. It makes no empirical claim about attribution or editorial judgment.

Automatic evaluation supplies several imperfect signals. ROUGE measures n-gram overlap with a reference summary [3]; BERTScore compares contextual token representations [4]. Purpose-built consistency evaluators include the classifier-based FactCC [5], question-answering approaches such as QAGS [6], FEQA [7], and QuestEval [8], and the NLI-based SummaC method [9]. Benchmarking and annotation studies including FRANK [10], fine-grained factuality annotation [11], and SummEval [12] show that metric behaviour depends on the label definition, dataset, and evaluation unit. Generation-based scoring such as BARTScore [13] and cross-dataset meta-evaluation in TRUE [14] further demonstrate that no single automatic signal covers every factual-consistency setting. Combining signals might recover complementary information, but a fitted combination can also add noise or overfit properties of the released systems. Improvement must therefore be tested on unseen source articles, not merely on randomly held-out summary rows from the same articles.

This study contributes a locked, reproducible head-to-head evaluation of a deliberately simple five-score ridge fusion against the published entailment score. Its technical emphasis is the evaluation design: grouping all summaries of an article in the same split, preserving a single primary contrast, and carrying article clusters through uncertainty estimation. The exercise also provides a realistic negative result. A model with more inputs is not presumed to be better; the conclusion follows the recorded held-out comparison. All features, human targets, and generated summaries originate in an existing public dataset. The only model fitted here is the lightweight metric-fusion regression.

2. Materials and Methods

2.1 Data Source and Scope

We used the official Google Research XSum hallucination-annotations repository at the pinned commit [15]. Its README states that the dataset is released under Creative Commons Attribution 4.0 and describes three human judgments per document–system pair. We downloaded the original eval_scores_xsum_summaries.csv, annotation files, and README; retained their SHA-256 hashes; and did not download or reproduce the underlying BBC article text. The experiment analyses the score CSV, whose key is a summarization-system name joined to a BBC article ID. The source README describes a larger 500-by-five annotation design, whereas the actual score file at the pinned commit has 498 article IDs and four system configurations. We recorded that release-specific discrepancy before viewing outcomes and used the observed file, not the nominal README count, as the study population.

The file provides five automatic inputs: ROUGE-1 F1, ROUGE-2 F1, ROUGE-L F1, BERTScore, and entailment probability. It also provides two human-derived outcomes. Faithful is the average proportion of summary words outside spans annotated as unsupported by the article, using the three released judgments. Factual is the mean of three binary factuality judgments and can therefore take values 0, one-third, two-thirds, or 1 [15]. These operational definitions are narrower than a newsroom assessment of accuracy across independent external sources. The precomputed automatic scores are used exactly as released; the original metric models were not run locally, and their scoring code and upstream model settings were not reconstructed in this experiment.

Each score row represents an article–generator pair, not a unique article or an independent human rater. The human judgments entered the released target columns before our analysis. We used those columns as published instead of recomputing them

from the annotation CSVs, because recomputation could silently change tokenization, span boundaries, or handling of missing worker records. That choice makes the target definition traceable to the publisher's own preprocessing, while retaining a clear boundary between our local computation and theirs. It also means that any annotation or preprocessing error in the original release carries into this reanalysis. The article IDs were treated as opaque grouping keys; the analysis did not require access to the BBC stories themselves.

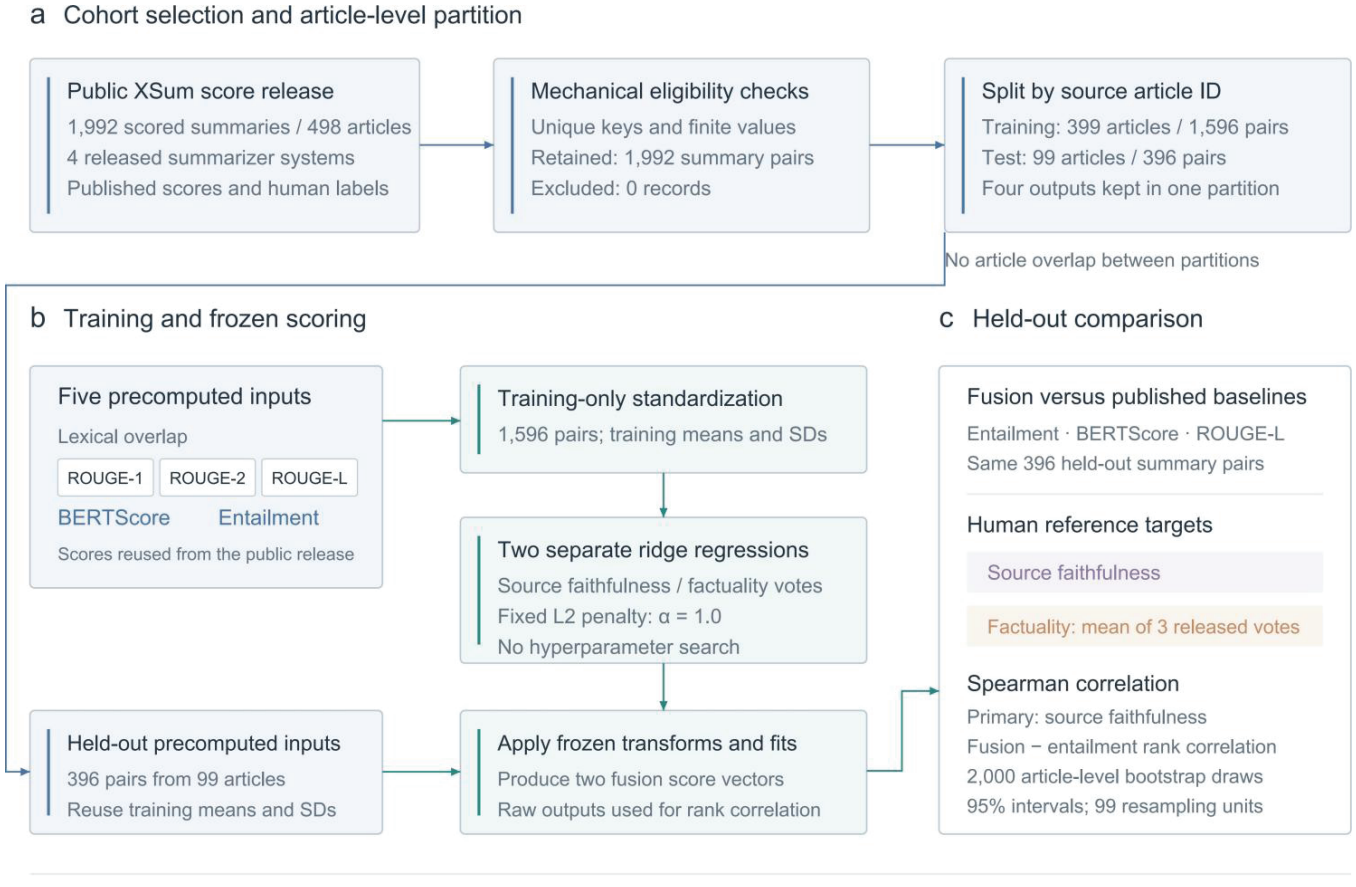
2.2 Frozen Eligibility and Article-Level Partition

The protocol was written and hashed before numerical outcome inspection. A record was eligible if its system–article key parsed unambiguously, its key was unique, and all five predictors and both targets were finite. No record was filtered by score magnitude, model family, or human label. We parsed the final underscore-delimited token of system_bbcid as the article ID, preserving the preceding tokens as the system identifier. The eligibility audit found 1,992 rows, 498 distinct article IDs, four systems with 498 rows each, and zero exclusions. Table 1 gives the frozen units and roles.

Table 1: Dataset and analysis units.

Item	Observed count	Use or definition
Released score records	1992	System–article pairs in the pinned CSV
Eligible records	1992	Finite inputs and targets; unique parsed key
Training partition	399 articles / 1596 pairs	Ridge fit only
Held-out partition	99 articles / 396 pairs	All reported endpoints
Released generators	4 configurations	99 test pairs per system
Exclusions	0	Predefined mechanical checks

Figure 1 summarizes the frozen record flow from released scores to article-held-out comparison.



Distinct article IDs were sorted by SHA-256 of the UTF-8 string 20261010|<bbcid>. The first floor(20%) were assigned to the test set and the rest to training. This deterministic operation yielded 99 held-out articles and 399 training articles. Each article contributed four system summaries, so all four were kept together. The test IDs, train IDs, source hashes, and split seed are retained with the code. The split is a prospective analysis choice for this reanalysis, not the publisher's original train/test partition. A small mechanical pilot processed the first 20 training article IDs, checked finite arrays and successful fit/predict/file writing, and reported no performance statistic. The full experiment followed that successful pilot.

The deterministic hash order is independent of the visible human scores and automatic metric values. It prevents a researcher from choosing a favourable held-out set after seeing the labels and permits an exact reconstruction from the saved article IDs. The 80/20 proportion applies to unique articles, not rows, and floor rounding explains why 20% of 498 becomes 99 rather than 100. We did not stratify by generator, topic, label, or date. Because every eligible article has the same four system configurations, generator counts remain equal in both partitions without a balancing operation. Article grouping is essential: if a row-level split were used, different summaries of the same source could appear on both sides of a fitted comparison, obscuring the intended assessment on previously unseen articles.

Figure 1: Document-level workflow for the article-held-out evaluation. All 1,992 eligible score rows derive from 498 source article IDs; the four system summaries for each article remain in one partition.

2.3 Comparison Systems

The primary baseline is the released entailment probability. Published BERTScore and ROUGE-L scores are secondary single-score baselines. The locally fitted fusion uses all five published inputs, ordered as ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, and entailment. We fit separate ridge regressions for faithfulness and factuality. For each fit, predictor means and population standard deviations were estimated on training records only. Zero-variance inputs would be centered and assigned scale one. With an unpenalized intercept and a fixed L2 penalty of 1.0, coefficients were obtained by solving the normal equations; no hyperparameter search, early stopping, test-set refit, or threshold selection was performed. The implementation uses NumPy rather than a downloaded large language model. The fitted feature moments, intercepts, and coefficients are saved in JSON.

More explicitly, for a standardized training matrix X and a human target vector y, the five coefficients solve $(X^T X + I)\beta = X^T(y-\bar{y})$; a test prediction is $\bar{y} + z\beta$, where z uses training means and standard deviations. The identity penalty applies only to coefficients, not the intercept $\bar{y}$. The same fixed design is fitted once for each human target, so the secondary factuality fit does not borrow its outcome from the faithfulness fit. This minimal model was selected to test whether already available metric signals add linear information beyond entailment under a reproducible, low-computation procedure. It is not a new natural-language inference model and does not generate or edit any summary.

Unclipped fusion predictions were used for rank correlation because clipping can create artificial ties. Predicted values and published baseline scores were clipped to [0,1] only for descriptive mean absolute error (MAE). Such MAE values are not a fair superiority test: the fusion is trained directly on the human target's scale, while the three published scores enter without recalibration. This distinction was set in the protocol before the outcome was read.

2.4 Outcomes and Statistical Analysis

The primary endpoint is Spearman's rank correlation, ρ, between each score and human Faithful on the same 396 held-out system summaries. The prespecified contrast is ρ(fusion, Faithful) minus ρ(entailment, Faithful). We computed ranks with average ties. Correlation with Factual, single-score secondary baselines, clipped MAE, and correlations within each generator system were descriptive secondary analyses. Each system-specific subset contains one summary for each of the 99 held-out articles. We did not select a best system or score after inspecting results.

For uncertainty, we resampled the 99 held-out article IDs with replacement 2,000 times using NumPy random seed 20261010. Each sampled article brought all four of its system-summary records into a draw. We recomputed correlations and the paired fusion-minus-entailment contrast in each draw, then reported the 2.5th and 97.5th percentile limits. A replicate would have been discarded if a correlation was undefined because a target or score was constant; none was discarded. The intervals express sampling variability under this article-level resampling scheme; they are not a guarantee of transport to another

publisher, time period, or summarization model. We report no p-values or claim of statistical superiority.

The correlation measures ordering, not error frequency at a chosen operational threshold. A positive ρ says that higher scores tend to accompany higher human labels across these released pairs; it does not say that a particular high-scoring summary is safe to publish. We evaluated all methods against the identical rows to keep contrasts paired. The bootstrap refits neither the ridge model nor the publisher's upstream metrics; it estimates uncertainty conditional on the saved training fit and the observed article sample. A fuller assessment of model-development uncertainty would repeat the entire fitting and split process, which was outside the frozen protocol. We make that conditional scope explicit because a narrow interval from a pair-level bootstrap could otherwise appear more decisive than the available independent articles permit.

2.5 Reproducibility and Permitted Use

The frozen protocol, amendment log, pinned source commit, raw-file hash manifest, pilot audit, split membership, full predictions, fitted coefficients, bootstrap draws, summary metrics, plotting scripts, and editable SVG figures are part of the accompanying research folder. The raw annotation files remain attributable to the Google Research release and its CC BY 4.0 licence [15]. The analysis neither redistributes BBC article text nor asks new people to judge summaries. In particular, the three original ratings are existing third-party dataset annotations; they are not observations collected by the current author.

3. Results

3.1 Record Flow and Target Distributions

All 1,992 released score rows passed the locked mechanical eligibility checks. There were 1,596 training records and 396 held-out records, corresponding to 399 and 99 distinct article IDs respectively. Figure 1 shows these counts and the comparison path. No held-out article ID appeared in training. The held-out set contains 99 summaries from each of the four released systems. Because the publisher's narrative count and score-file count differ, these empirical denominators refer specifically to the pinned eval_scores_xsum_summaries.csv release.

Human labels were not uniformly distributed (Figure 2). The 396 held-out faithfulness scores span [0,1], with 95 in [0,0.2), 40 in [0.2,0.4), 56 in [0.4,0.6), 84 in [0.6,0.8), and 121 in [0.8,1.0]. The last interval includes 1. The median faithfulness score was 0.613. For factuality, 313 of 396 records (79.0%) had a mean human vote of zero; 10 (2.5%) had one-third, six (1.5%) had two-thirds, and 67 (16.9%) had one. The concentration at zero makes the factuality correlation sensitive to ties and limits the interpretability of a small numerical difference.

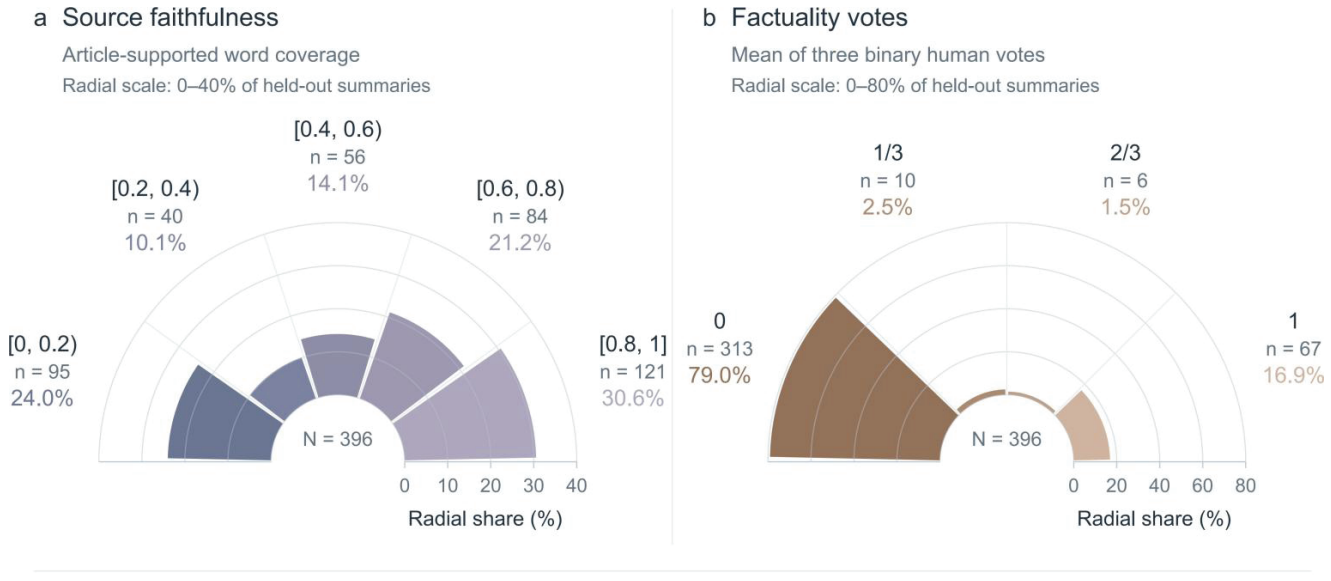
The two targets have different measurement granularity. Faithfulness can take many fractions because it aggregates judgments over summary words as well as annotators. Factuality has only four possible averaged-vote levels. The observed factuality distribution is therefore not a histogram of independently verified events, and its zero category must not be labelled a 79% rate of objectively false news summaries. It is the share of these released generated-summary records for which all three recorded factuality votes were negative. Figure 2 and all subsequent denominators describe the held-out set, not the whole XSum corpus or contemporary news generation systems.

3.2 Prespecified Faithfulness Comparison

On the 99-article, 396-summary held-out set, the fixed fusion reached Spearman $\rho = 0.425$ (95% article-bootstrap interval 0.325–0.515). The published entailment score reached $\rho = 0.445$ (0.342–0.532). The prespecified paired difference was 0.020, with a 95% interval from 0.100 to 0.066 (Table 2 and Figures 3–4). The interval includes both a moderate disadvantage and a modest advantage for fusion. It does not support a claim that the five-score model improves rank agreement with the released human faithfulness labels. BERTScore and ROUGE-L correlations were lower at 0.160 and 0.157 respectively on this same test set; these are descriptive comparisons, not independent replications of the original metric papers.

The fitted faithfulness regression had a training-label intercept of 0.579. In the prespecified feature order, its coefficients on standardized inputs were 0.077, 0.040, 0.031, 0.050, and 0.119. These coefficients document the local fit and are not causal effects of changing a metric. The modest gain in some generator-specific subsets and losses in others (Table 3) are consistent with heterogeneous alignment between the five inputs and human labels. With only 99 records per subset and no frozen subgroup contrast, the table is diagnostic rather than a basis for declaring one system superior.

Figure 2: Released human labels for 396 held-out summaries. Radial length encodes share, with scales of 0–40% for source faithfulness (a) and 0–80% for factuality votes (b).



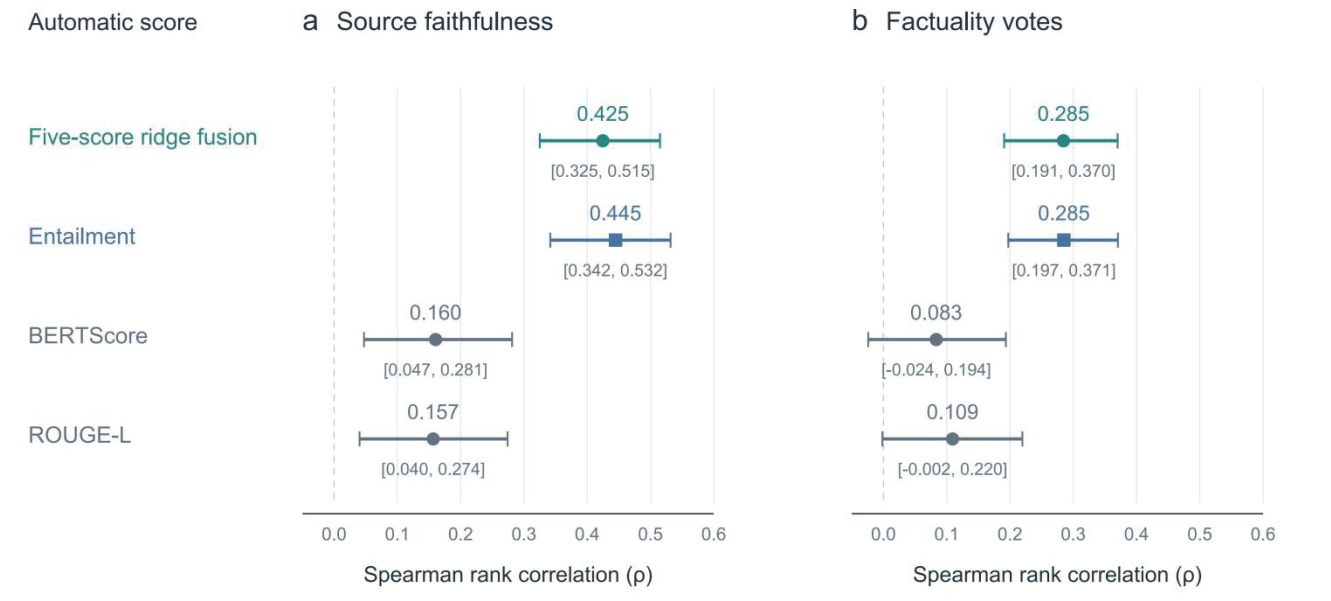
Held-out sample: 396 summaries from 99 articles; human labels are reused from the public XSum release.
Radial length encodes share. Panel a: 0–40%; panel b: 0–80%. Compare the printed percentages across panels.
Faithfulness bins include the lower bound; the final bin also includes 1. Values are rounded to one decimal place.

Table 2: Held-out Spearman correlation with human source faithfulness (396 pairs from 99 articles).

Score	Spearman ρ	95% article-bootstrap interval
Fusion	0.425	0.325 to 0.515
Entailment	0.445	0.342 to 0.532
BERTScore	0.160	0.047 to 0.281
RL	0.157	0.040 to 0.274

Paired fusion minus entailment: -0.020 (-0.100 to 0.066); 2,000 article-level bootstrap draws; 0 undefined.

Figure 3: Held-out Spearman correlations with the two human targets. Points are estimates; horizontal intervals are 95% percentiles from 2,000 article-level bootstrap draws.



Points: estimates. Lines and brackets: 95% intervals from 2,000 article-level bootstrap draws; n = 99 articles / 396 pairs.

The overall correlation pools summaries generated by all four systems. A pooled association can include between-system

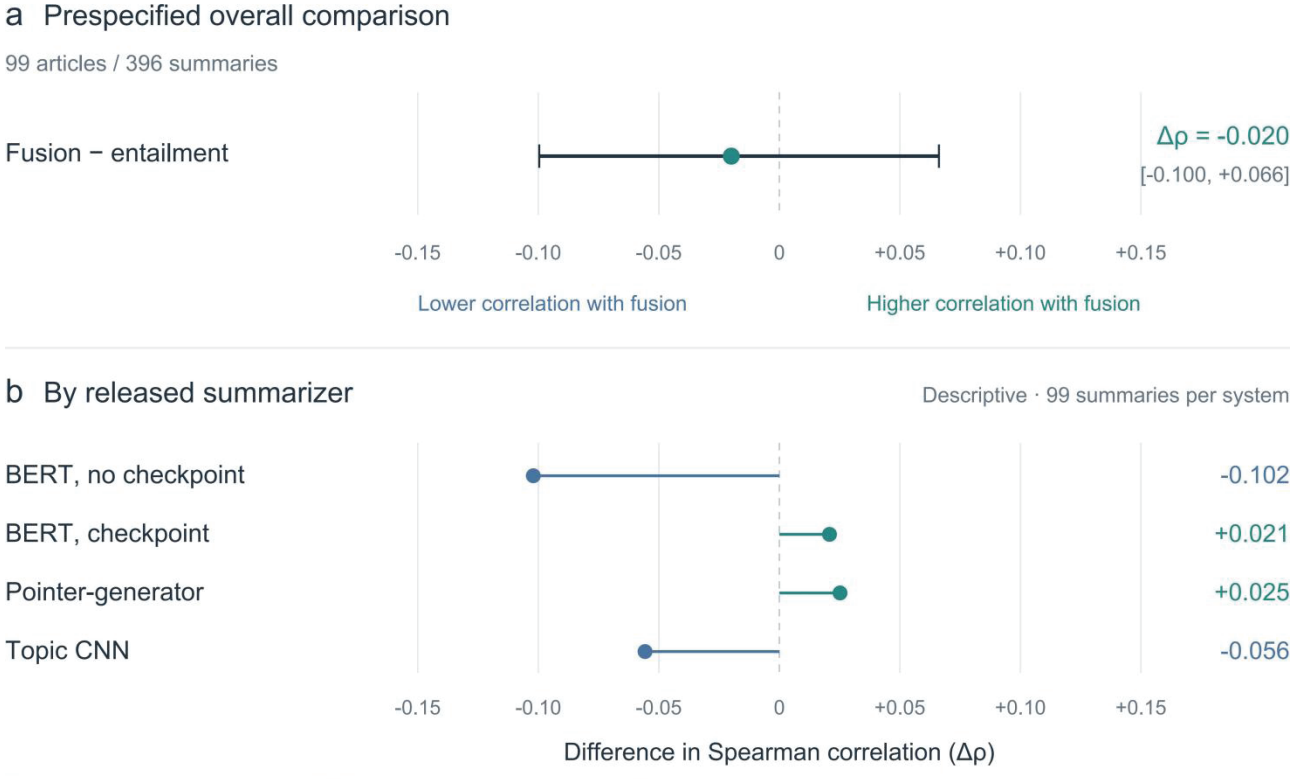
differences in typical metric levels as well as variation among articles within a system. Table 3 separates the systems descriptively, without attempting to estimate a hierarchical model. Fusion was lower than entailment by 0.102 in the no-checkpoint BERT group and 0.056 in the topic-CNN group, while it was higher by 0.021 and 0.025 in the other two groups. These signs show why the main result should be read as an average over the exact released mix rather than a universal ranking for every summarizer. The subgroup values have no individual confidence intervals in the frozen analysis and are not used to revise the primary comparison.

Table 3: Descriptive faithfulness correlations by released generator.

Generator	Pairs	Fusion ρ	Entailment ρ	Difference
BERT, no checkpoint	99	0.287	0.389	-0.102
BERT, checkpoint	99	0.539	0.518	0.021
Pointer-generator	99	0.487	0.461	0.025
Topic CNN	99	0.329	0.385	-0.056

System names follow the release; differences were not prespecified subgroup tests.

Figure 4: Overall and descriptive generator-specific differences in source-faithfulness correlation. Positive values favour fusion; the prespecified overall interval crosses zero.



Only the overall comparison has a 95% bootstrap interval; subgroup differences have no separate uncertainty intervals.

3.3 Secondary Factuality Target

Against the released factuality vote, fusion produced ρ = 0.285 (95% interval 0.191–0.370) and entailment ρ = 0.285 (0.197–0.371). Their unrounded difference was 0.0007, with an article-bootstrap interval of 0.081 to 0.087. BERTScore and ROUGE-L yielded 0.083 and 0.109 respectively (Table 4). This secondary result is close to a tie between fusion and entailment in the current split. It must not be interpreted as a direct comparison with external truth, because the outcome is the publisher's existing three-person mean of yes/no judgments on the selected generated summaries.

The factuality ridge intercept was 0.159; standardized-input coefficients were 0.016, 0.007, 0.022, 0.014, and 0.086 in the same five-feature order. These small-scale values partly reflect a target concentrated at zero. The score-derived predicted values and target labels are preserved row by row, so alternative calibration analyses can be conducted without rerunning the original annotations. We do not present a newly chosen binary decision threshold or classification accuracy, because neither was in the frozen protocol.

Table 4: Held-out Spearman correlation with human factuality vote (396 pairs from 99 articles).

Score	Spearman ρ	95% article-bootstrap interval
Fusion	0.285	0.191 to 0.370
Entailment	0.285	0.197 to 0.371
BERTScore	0.083	-0.024 to 0.194
RL	0.109	-0.002 to 0.220

Paired fusion minus entailment: -0.001 (-0.081 to 0.087); 2,000 article-level bootstrap draws; 0 undefined.

3.4 Descriptive Absolute Error

After clipping scores to [0,1], the faithfulness MAE was 0.261 for trained fusion and 0.378 for raw entailment; the factuality MAE values were 0.257 and 0.329. Table 5 includes the other raw baseline scores. These numbers describe distances on the publisher's label scale. They cannot establish a fair calibration advantage of score fusion, because the trained model was allowed to learn the outcome scale and the published baselines were not. The prespecified ranking contrast remains the appropriate basis for the main conclusion.

Table 5: Descriptive clipped mean absolute error on the held-out labels.

Score	Faithfulness MAE	Factuality MAE
Fusion	0.261	0.257
Entailment	0.378	0.329
BERTScore	0.305	0.646
RL	0.387	0.326

Fusion was trained on the label scale; uncalibrated published baselines were not. MAE is descriptive only.

Table 6 gives the saved training moments and standardized coefficients for both local fits. All five input standard deviations were nonzero in the training partition, so the zero-variance safeguard was not activated. The negative ROUGE-2 and ROUGE-L coefficients in the faithfulness model should not be read as evidence that overlap intrinsically harms faithfulness: these are conditional weights in a correlated five-input regression. Their magnitude and sign depend on the training partition, feature scaling, regularization, and the other scores included in the model. The table is supplied for reproducibility and model inspection, not for a causal explanation of individual summaries.

Table 6: Saved ridge inputs and standardized coefficients from the 1,596 training pairs.

Input	Train mean	Train SD	Faithful β	Factual β
R1	0.336	0.151	0.077	0.016
R2	0.125	0.131	-0.040	-0.007
RL	0.270	0.143	-0.031	0.022
BERTScore	0.762	0.060	0.050	0.014
Entailment	0.357	0.460	0.119	0.086

Intercepts: Faithful 0.579; Factual 0.159. Fixed ridge penalty = 1.0.

3.5 Integrity Checks and Analytical Boundaries

The audit found no duplicate or malformed keys, missing outcomes, or nonfinite inputs. Saved held-out predictions contain each eligible test key once; train and test article IDs are disjoint. The bootstrap file preserves every numerical draw, permitting independent recomputation of the intervals. The 20-article pilot logged finite fit and prediction values only. No failed model call or bootstrap draw needed post hoc repair.

These mechanical checks support regeneration of the tables but do not validate the original annotations against BBC evidence. "Held-out" means article IDs withheld from our ridge fit; it does not mean a new publisher, year, or newsroom setting.

4. Discussion

The central finding is a non-improvement under a specific and auditable test. Combining overlap, contextual similarity, and entailment into a fixed ridge model did not yield a higher held-out rank correlation with human source-faithfulness scores than entailment alone. The point estimate slightly favoured entailment, while the cluster-bootstrap interval was too broad to rule out either a modest benefit or a larger loss. The study does not demonstrate equivalence: lack of a detected advantage is not proof that all future fusions will fail. It limits only the evaluated feature set, model configuration, label construction, four released summarizers, and article-level split.

The result is technically informative because adding correlated evaluation metrics can improve in-sample fit without supplying a robust new signal. ROUGE and BERTScore often reward lexical or semantic resemblance, but resemblance does not itself establish that an assertion is supported by the article. In the present data, the entailment score already has stronger marginal rank alignment with source faithfulness than the two reported similarity baselines. The fused model still uses entailment and has a positive standardized coefficient for it, yet additional inputs did not improve the primary ranking. This interpretation follows the actual held-out correlations and coefficient record; it is not a mechanistic proof of why the model behaved this way.

The article-grouped design is important for the stated generalisation question. A random split of system-summary pairs could place one BBC article in both fitting and testing, allowing article-specific properties to affect both sides. Our hash partition kept the four summaries attached to each article together. The bootstrap likewise resampled articles, retaining within-article dependence. Although the analysis has 396 test rows, its independent resampling units are 99 source articles. The interval therefore reflects less information than a naïve row-level analysis would imply. It does not incorporate uncertainty in the original three human judgments or uncertainty in the publisher's precomputed model scores.

Document grouping does not eliminate every dependency: the four systems share article sources and possibly upstream preprocessing. The bootstrap retains within-article dependence but treats distinct article IDs as exchangeable. Unmeasured topic or publication-period clustering could change uncertainty, so the interval is specific to the frozen article-level scheme.

The two human targets must also remain distinct. The span-based Faithful score is specifically article-supported word coverage. It may miss a misleading emphasis that uses individually supported words, and it is not an expert editorial review. The Factual vote asks whether a hallucinated summary is factual, based on the existing released labels, and is heavily tied at zero in the held-out subset. Neither outcome includes explicit claim–source attribution, quote provenance, framing preservation, selection of salient facts, or newsroom policy compliance. The earlier proposed three-dimensional study cannot be completed from these labels. Future research on those dimensions would require a codebook, licensed article evidence, independent qualified raters, adjudication, and institutional review as applicable; none of that has been represented as completed here.

Several other limitations constrain external validity. XSum consists of BBC news, and these released outputs are from four historical summarizer configurations. Contemporary large language models, multilingual stories, live updates, and longer multi-document summaries may yield different score and error distributions. We relied on the publisher's precomputed metrics rather than rerunning each upstream evaluator; this keeps the local experiment feasible and exactly tied to the release but prevents diagnosis of upstream metric-version choices. We performed one prespecified split and one ridge penalty, without repeated split validation or hyperparameter tuning. The system-specific patterns are descriptive and could shift with a different article sample. Finally, the raw annotation files have a different structure from the score table, so we did not assert a joined intrinsic/extrinsic error analysis without an unambiguous audit trail.

There is also a measurement limit to the broad phrase "fact checking." The human faithfulness label grades support within a particular source article; it does not independently verify a real-world claim against primary documents, witnesses, or later corrections. The factuality vote is likewise the released annotators' judgment, not an institutional verification of the underlying BBC report. A newsroom may care about additional harms that neither label captures, including misquoted speakers, missing caveats, stale numbers, or an answer that overstates certainty while reusing source vocabulary. Such cases need claim-level provenance and editorial review; they cannot be inferred from this score table. The absence of those labels

drove the narrowed research question before the results were used for a paper claim.

For practical deployment, a moderate global association does not specify a safe cut-off. Threshold choice would depend on the frequency and consequence of unsupported claims in the actual workflow, and it would require calibrated probabilities or decision costs validated on that workflow's articles. The present experiment has neither an operating threshold nor a prospective newsroom sample. Its contribution is instead a reproducible comparison that can prevent an unjustified claim that simply combining fashionable metrics improves reliability. A follow-up could test a separate, licensed contemporary corpus, predefine claim-level outcomes with qualified raters, and compare calibrated baselines under article and publisher shifts. That follow-up would be a new study, not a result of this one.

The work illustrates a useful decision rule for engineering practice: a more elaborate summary-quality score should earn its complexity on a held-out, human-labelled unit appropriate to deployment. In this case, the simple entailment baseline remains at least as persuasive for ranking the released faithfulness labels. Neither score should replace a human source check for a publishable news item, because moderate correlation across a dataset does not certify an individual summary. Any future operational threshold would require a separate validation design measuring error costs at the intended newsroom prevalence.

5. Conclusion

A reproducible article-held-out reanalysis of 1,992 released news-summary records found no evidence that fixed five-score ridge fusion improves agreement with existing human source-faithfulness labels over the published entailment score. The locked test set comprised 99 BBC article IDs and 396 system summaries. The primary correlation difference was 0.020 with a 95% article-bootstrap interval of 0.100 to 0.066. A secondary human factuality-vote analysis likewise showed no meaningful numerical advantage. The findings apply to the released label definitions and system outputs; source attribution and editorial fidelity remain untested.

Funding

No

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 1797–1807. https://doi.org/10.18653/v1/D18-1206

[2] Maynez, J., Narayan, S., Bohnet, B., et al. (2020). On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

[3] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out, 74–81. https://aclanthology.org/W04-1013/

[4] Zhang, T., Kishore, V., Wu, F., et al. (2020). BERTScore: Evaluating Text Generation with BERT. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1904.09675

[5] Krys´cin´ski, W., McCann, B., Xiong, C., et al. (2020). Evaluating the Factual Consistency of Abstractive Text Summarization. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 9332–9346. https://doi.org/10.1 8653/v1/2020.emnlp-main.750

[6] Wang, A., Cho, K., & Lewis, M. (2020). Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5008–5020. https://doi.org/10.18653/v1/2020.acl-main.450

[7] Durmus, E., He, H., & Diab, M. (2020). FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational

Linguistics, 5055–5070. https://doi.org/10.18653/v1/2020.acl-main.454

[8] Scialom, T., Dray, P.-A., Lamprier, S., et al. (2021). QuestEval: Summarization Asks for Fact-based Evaluation. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 6594–6604. https://doi.org/10.18653/v1/20 21.emnlp-main.529

[9] Laban, P., Schnabel, T., Bennett, P. N., et al. (2022). SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. Transactions of the Association for Computational Linguistics, 10, 163–177. https://doi.org/10.1162/tacl_a_00453

[10] Pagnoni, A., Balachandran, V., & Tsvetkov, Y. (2021). Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4812–4829. https://doi.org/10.18653/v1/2021.naacl-main. 383

[11] Goyal, T., & Durrett, G. (2021). Annotating and Modeling Fine-grained Factuality in Summarization. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1449–1462. https://doi.org/10.18653/v1/2021.naacl-main.114

[12] Fabbri, A. R., Krys´cin´ski, W., McCann, B., et al. (2021). SummEval: Re-evaluating Summarization Evaluation. Transactions of the Association for Computational Linguistics, 9, 391–409. https://doi.org/10.1162/tacl_a_00373

[13] Yuan, W., Neubig, G., & Liu, P. (2021). BARTScore: Evaluating Generated Text as Text Generation. Advances in Neural Information Processing Systems, 34, 27263–27277. https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1 a62fee3d9dd-Abstract.html

[14] Honovich, O., Aharoni, R., Herzig, J., et al. (2022). TRUE: Re-evaluating Factual Consistency Evaluation. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3905–3920. https://doi.org/10.18653/v1/2022.naacl-main.287

[15] Google Research Datasets. (n.d.). XSum hallucination annotations (commit 4abdbb2c807cb93c1df1188a1992d20a55cdcc34) [Data set]. GitHub. Retrieved October 10, 2026, from https://github.com/google-research-datasets/xsum_hallucination_ann otations