Enhancing Factuality through Consensus and Consistency in Summarization Using Minimum Bayes Risk Decoding

Riza Setiawan Soetedjo[1], Yusuke Sakai[1], Hidetaka Kamigaito[1], Jingun Kwon[2], Manabu Okumura[3], Taro Watanabe[1]
[1]Nara Institute of Science and Technology (NAIST) [2]Chungnam National University
[3]Institute of Science Tokyo
riza.setiawan_soetedjo.rs6@naist.ac.jp
{sakai.yusuke.sr9, kamigaito.h, taro.watanabe}@is.naist.jp
jingun.kwon@cnu.ac.kr, oku@first.iir.isct.ac.jp

Abstract

Improving the quality of model-generated summaries, especially factuality, the accuracy of a summary with respect to its source content, remains a challenge. While reranking could select the optimal output from multiple generated candidates, it is limited to only using the source as guidance, resulting in unreliable summaries. To address this limitation, we propose ConSUM that reranks candidate summaries by considering two factors: *consistency* to the source document and *consensus* among the other candidates. Consensus is established using Minimum Bayes Risk (MBR) decoding over the set of generated summaries, while ensuring consistency by employing factuality-aware metrics that compare the summary against the source. Rigorous testing demonstrates that our system is competitive with existing methods, with human evaluations further confirming that its generated summaries are preferred over those from other systems. Our code is available at https://github.com/naist-nlp/ConSUM.

1 Introduction

Document Summarization is a task to summarize a lengthy document while retaining its most important information. Thus, to evaluate a generated summary, we should focus on its factuality, i.e., how the generated summary aligns with the original document, as well as commonly used metrics in Natural Language Generation (NLG), such as fluency and coherence (Fabbri et al., 2021).

Reranking is an established method to improve summary quality, including factuality (Sul and Choi, 2023; Ravaut et al., 2022), which involves generating multiple candidate summaries and ranking them using specific metrics reflecting summarization quality. Since we cannot use gold references, reference-free metrics are used to rerank the candidates by using, e.g., the source document as their ground truth (Dixit et al., 2023).

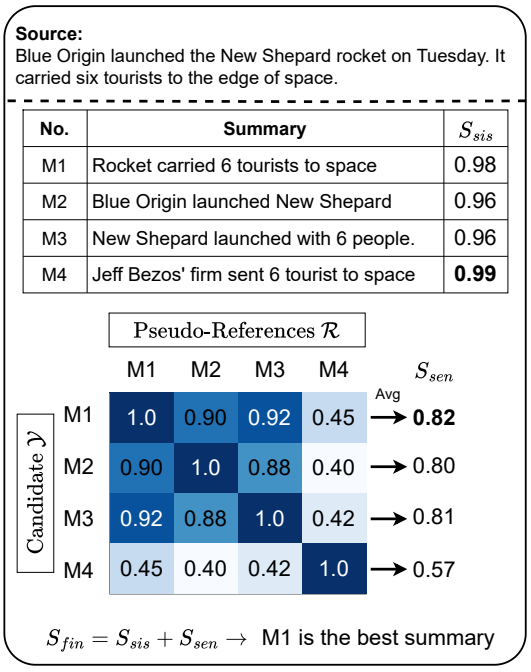

Figure 1: Case where reference-free metric resulted in factual inconsistency. S_{sis} is consistency score to the source document; S_{sen} is consensus score over among candidate summaries; S_{fin} is the combined score.

Even though the success, the current reranking in summarization has some limitations: First, it cannot use reference-based metrics that correlate well to human evaluation due to requiring gold references; Second, relying only on the source can result in factual inconsistencies by failing to penalize specific errors due to the broad scoring nature of reference-free metrics, as shown in Figure 1; Third, relying only on a single metric may cause overfitting to that metric and inherit its bias.

To address these problems, we propose a novel reranking method, **Con**sistency and **Con**sensus in **Sum**marization (ConSUM), that combines two factors: the *consensus* among candidate summaries and the *consistency* of each summary to the source

Findings of the Association for Computational Linguistics: ACL 2026, pages 41686–41713

document for factual summary generation. Inspired by Minimum Bayes Risk (MBR) decoding (Eikema and Aziz, 2020), we achieve consensus by identifying the most representative summary among the generated candidates, relying on a reference-based factuality metric using pseudo-references sampled from the same model to serve as proxies for the ground truth. Consistency for the source document, meanwhile, is measured using reference-free metrics. Our method operates in a simple three-step process: (1) **Generate** multiple candidate summaries and pseudo-references given the source text; (2) **Score** each candidate using both MBR decoding and reference-free metric; and (3) **Rerank** each candidate by combining both scores and select the one with the highest final score.

We benchmarked our method on the CNN/DailyMail (Nallapati et al., 2016) and XSum (Narayan et al., 2018) datasets using three Pretrained Language Models (PLMs) and a Large Language Model (LLM), evaluated in two categories of metrics: Quality, e.g., ROUGE (Lin, 2004) and Factuality, e.g, MENLI (Chen and Eger, 2023). Furthermore, we did extensive experiments to find the optimal hyperparameters in ConSUM and their influence on summarization quality. Finally, we conducted a human evaluation to assess our method qualitatively. Experimental results consistently demonstrate that our method achieves superior factuality scores across both automatic and human evaluations. This shows that combining *consensus* and *consistency* succeeded in improving the factuality of a summary.

2 Related Work

Improving Factuality in Summary Current approaches to improving factuality often use existing evaluation metrics to improve the summarization model. Research in this area typically either specifies only improving factuality (Liu et al., 2023) or attempts to balance multiple aspects simultaneously (Ryu et al., 2024; Dixit et al., 2023). However, these methods face two limitations: requiring high-quality gold references, or relying solely on the source document for guidance. This reliance on a single reference signal arises because creating human-written "gold" summaries for comparison is notoriously labor-intensive. On the other hand, only using the source document is unreliable.

Reranking in Summarization Reranking in summarization is a two-step approach: (1) a summarization model generates multiple summaries based on the source document; (2) a reranker model reranks the summaries based on a specific evaluation metric. Liu and Liu (2021) trains a reranker model using ROUGE, while Dixit et al. (2023) trains a reranker model using FactCC and ROUGE to balance between factuality and quality. As an alternative to train a supervised reranker model, some studies have proposed methods to rerank in an unsupervised setup (Ravaut et al., 2023; Suzgun et al., 2023). However, they only use quality metrics such as ROUGE and BERTScore.

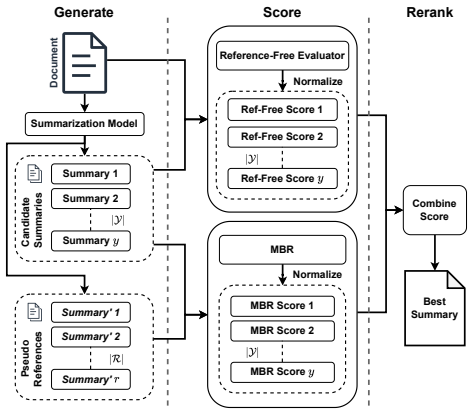

Figure 2: Overview of ConSUM comprising three steps: Generate, Score, and Rerank.

Minimum Bayes Risk (MBR) Decoding MBR decoding was used in Neural Machine Translation (NMT) to address the mode-seeking problem inherent in standard Maximum a Posteriori (MAP) decoding (Eikema and Aziz, 2020). Instead of selecting a single most probable output, MBR first generates a pool of candidate texts, then selects the candidate that maximizes an expected utility function when scored against a set of references, which are usually generated by the same model (Ohashi et al., 2024). The utility function itself is typically an existing evaluation metric. While studies show that MBR effectively optimizes for whichever metric is chosen as the utility function (Bertsch et al., 2023), this can lead to a form of reward hacking or metric bias. Although this issue has been extensively researched in NMT (Kovacs et al., 2024; Müller and Sennrich, 2021), its impact within the summarization task remains largely unexplored.

3 Proposed Method: ConSUM

The overview of ConSUM is presented in Figure 2. It comprises three steps: (1) *Generate Candi-*

dates and Pseudo-references – The summarization model generates multiple candidate and pseudo-references summaries given the source document; (2) *Score Candidates* – Each candidate is scored using two distinct evaluators, one assessing its consistency to the source document and the other assessing the consensus between the candidates and the pseudo-references; (3) *Rerank Candidates* – Both scores are combined and the candidate with the highest overall score is selected as the final output.

3.1 Generate Candidates and Pseudo-references

To enhance factuality, we distinguish between the *candidates* $\mathcal{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{|\mathcal{Y}|}\}$ (the pool of potential outputs) and the *pseudo-references* $\mathcal{R} = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_{|\mathcal{R}|}\}$ (proxies for the ground truth). Both are sampled from the same model θ given a source s, but they serve distinct roles and may use different decoding strategies, denoted as θ_{cand} and θ_{ref}:

$$\mathcal{Y} \sim p(\boldsymbol{y}|s;\theta_{\text{cand}}) \quad \text{and} \quad \mathcal{R} \sim p(\boldsymbol{r}|s;\theta_{\text{ref}}). \quad (1)$$

While prior work often sets $\mathcal{Y} = \mathcal{R}$ (Suzgun et al., 2023), we treat them as distinct sets. This allows us to optimize $\mathcal{R}$ to better capture the model's true distribution for consensus estimation (Kamigaito et al., 2025), independent of the strategy used to generate diverse *candidates* $\mathcal{Y}$ (see Appendix A.2).

3.2 Score Candidates

We assess the validity of a candidate y using two signals: its *consistency* with the source document and its *consensus* with the pseudo-references $\mathcal{R}$. Consistency-based scores have been used widely to improve summarization (Liu and Liu, 2021; Dixit et al., 2023). However, we posit that relying solely on the source is insufficient due to metric bias and a lack of granularity, as illustrated in Figure 1. Since high-quality gold references are often inaccessible, we leverage pseudo-references $\mathcal{R}$ as a proxy reference signal, used in conjunction with the source document. This allows us to mitigate counterfactual outliers by penalizing information that diverges from the model's consensus distribution.

Consistency-based Scoring We define the consistency between the source document $\boldsymbol{s}$ and its candidate $\boldsymbol{y}_i \in \mathcal{Y}$, generated in §3.1, as:

$$S_{sis}(\boldsymbol{y}_i, \boldsymbol{s}) = FM(\boldsymbol{y}_i, \boldsymbol{s}), \quad (2)$$

where $FM(\boldsymbol{y}_i, \boldsymbol{s})$ is a reference-free factuality metric. In this study, we use FENICE (Scirè et al., 2024) and FIZZ (Yang et al., 2024) as the reference-free metric. We chose the metrics because they represent the state-of-the-art (SOTA) factuality-based reference-free metrics for summarization task[1]. Both metrics operate on a similar principle (see Appendix B).

Consensus-based Scoring We measure the consensus between a candidate $\boldsymbol{y}_i \in \mathcal{Y}$ and pseudo-references $\mathcal{R}$ by incorporating MBR decoding as follows:

$$S_{sen}(\boldsymbol{y}_i, \mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum_{j=1}^{|\mathcal{R}|} u(\boldsymbol{y}_i, \boldsymbol{r}_j), \quad (3)$$

where $u(\boldsymbol{y}_i, \boldsymbol{r}_j)$ is a utility function to calculate the validity of each candidate $\boldsymbol{y}_i$ using each pseudo-reference $\boldsymbol{r}_j$. In this study, we specifically choose MENLI (Chen and Eger, 2023) as the utility function, as it is a SOTA NLI-based metric designed specifically for summarization. The metric aligns with our goal of "consensus" driven by factual agreement, rather than just lexical or semantic popularity which are measured through ROUGE and BERTScore (Ravaut et al., 2023; Suzgun et al., 2023).

3.3 Rerank Candidates

To ensure a balanced contribution, we normalize and combine S_{sis} and S_{sen} as follows:

$$\begin{aligned} &S_{fin}(\boldsymbol{y}_i, \boldsymbol{s}, \mathcal{R}) \\ =&wZ(S_{sen}(\boldsymbol{y}_i, \mathcal{R}))+(1-w)Z(S_{sis}(\boldsymbol{y}_i, \boldsymbol{s})). \end{aligned} \quad (4)$$

Z indicates z-score normalization and w $(0 \leq w \leq 1)$ is a hyperparameter to adjust the importance of S_{sen}, where 0 and 1 means the scoring only uses S_{sis} and S_{sen}, respectively. Finally, we choose the best candidate $\hat{\boldsymbol{y}}$ based on S_{final} as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} S_{fin}(\boldsymbol{y}, \boldsymbol{s}, \mathcal{R}). \quad (5)$$

Unlike prior work relying solely on reference-free metrics (Dixit et al., 2023) or limiting MBR to only use BERTScore (Suzgun et al., 2023), we combine both paradigms. We identify the best candidate by aggregating its *consensus* score (using MENLI against the pseudo-references $\mathcal{R}$) and its *consistency* score (using FENICE or FIZZ against the source document).

[1] as of January 2025

Group	Reference-Based	Reference-Free
Quality	ROUGE, BERTScore, MoverScore	—
Factuality	MENLI	UniEval, FENICE, FIZZ, SimCLS

Table 1: Metrics by group and type by reference.

4 Experiments

We test our method on two news summarization datasets, using fine-tuned PLMs and a LLM as the summarization models, evaluated on metrics from two aspects: Quality and Factuality.

4.1 Experimental Settings

Datasets We evaluated our method on two widely-used news summarization datasets: CNN/DailyMail (CNN/DM) (Nallapati et al., 2016) and XSum (Narayan et al., 2018). CNN/DM is characterized by its relatively extractive summaries, where summary sentences are often copied directly from the source article. In contrast, XSum is known for its highly abstractive, single-sentence summaries that condense the main point of an article. Although these are standard benchmarks, recent work (Zhang et al., 2024) has highlighted the low quality of their "gold" references. These noise allow us to demonstrate that ConSUM can improve factuality through the model's consensus and consistency with the source document.

Models We used three PLMs, i.e, BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020), and T5 (Raffel et al., 2023), which were fine-tuned to the respective datasets. In addition, we experimented Llama-3 (Grattafiori et al., 2024) to test our method with LLMs using nucleus sampling (Holtzman et al., 2020)[2]. Despite generally outperforming PLMs, LLMs exhibit comparable summarization performance due to hallucinations (Chhabra et al., 2024), motivating us to evaluate both in this study. Furthermore, we explored two types of decoding for the PLMs: Diverse Beam Search (DBS) (Vijayakumar et al., 2018) and Epsilon Sampling (Hewitt et al., 2022). For DBS, we implemented the setting from Liu and Liu (2021) and refer to it as `beam-sim` in this study. On the other hand, we implemented the best epsilon setting from NMT (Freitag et al., 2023) and refer to it as `epsilon` in this study. We also explored alternative settings, as discussed in Appendix D.

Metrics Generated summaries are assessed on two metric groups: Quality and Factuality, as shown in Table 1. We used three reference-based metrics in the Quality group: ROUGE (Lin, 2004), BERTScore (Zhang* et al., 2020), and MoverScore (Zhao et al., 2019). In Factuality group, we utilized three scores provided by MENLI (Chen and Eger, 2023), i.e. entailment (EM), contradiction (CM), and summarization (SM) as the reference-based metric, UniEval (Zhong et al., 2022), FENICE (Scirè et al., 2024), and FIZZ (Yang et al., 2024) as the reference-free metrics. Quality metrics are the standard benchmarks in summarization studies. We also included UniEval to measure multiple dimensions alongside factuality (i.e. fluency and coherence). Although SimCLS underperformed as a reranking component (see §4.2), we retained it as a strong reference-free evaluator to verify our results against a finetuned model. Statistical significance is assessed using paired-bootstrap resampling (Koehn, 2004) with 10,000 iterations. We established a significance level of $p < 0.05$, applying the Bonferroni correction to account for multiple comparisons against the three baselines[3]. Implementation details are found in Appendix B.

Hyperparameters Our preliminary study (Appendix A) highlights the importance of using diverse and unbiased pseudo-reference sets, as well as maintaining the weight between consensus and consistency, to achieve optimal performance. We determined the best hyperparameters to be **16 candidates**; **64 pseudo-references, generated via epsilon sampling**; and **a combination weight (w) of 0.75**. The weight (w) means the contribution of S_{sen} and S_{sis} is 75:25 to the S_{final} score. In this study, our scorer methods are formatted as `Scorer-`w, where `Scorer` refers to either FENICE or FIZZ and w represents the weight assigned to the MBR score (S_{sen}). The setting with $w = 0.0$ uses only the named scorer (consistency-only), which serves as a baseline. On the other hand, $w = 1.0$ uses only the MBR score (consensus-only) and its `Scorer` is denoted as `MBR`. Conversely, a setting with $0 < w < 1$ indicates a linear combination of the named scorer and MBR score.

[2]Implementation details are in Appendix C.

[3]The significance test is done before the results are rounded.

	FENICE		FIZZ		SimCLS	
w	CNNDM	XSUM	CNNDM	XSUM	CNNDM	XSUM
0.00	81.05	34.67	14.15	17.37	12.76	8.62
0.25	62.29	**77.83**	36.88	13.23	24.51	15.60
0.50	72.64	75.95	49.76	26.86	39.57	26.87
0.75	**81.05**	77.52	**71.08**	55.03	63.51	55.06
1.00	68.86	39.70	69.69	**76.19**	**65.35**	**90.91**

Table 2: Average normalized of all metrics for CNN/DM and XSum datasets. The 'no mbr' and 'mbr only' settings are denoted by 0 and 1, respectively. Best scores for each reranker are in bold.

Baselines We evaluate our proposed configurations (MBR-1.0, FENICE-0.75, and FIZZ-0.75) against three baselines: `Baseline` (standard decoding without MBR) and the consistency-only rerankers (FENICE-0.0 and FIZZ-0.0).

4.2 Optimal Weight Combination (w)

Experiment Settings To leverage both *consensus* and *consistency*, our system combines MBR scores with reference-free metric scores. We investigated the optimal combination weight (w) using three distinct metrics: the factuality-focused FENICE (Scirè et al., 2024) and FIZZ (Yang et al., 2024), and SimCLS (Liu and Liu, 2021), which balances factuality and quality. We evaluated weights $w \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, spanning from only using reference-free scoring (0.0) to only using MBR (1.0). Experiments utilized the candidate samples generated using epsilon sampling for PLMs and nucleus sampling for LLM (see Appendix A.2). We fixed the pseudo-references using 64 samples generated via epsilon sampling.

Evaluation We evaluated the top-ranked summary selected by each weight and reranker combination using the two groups of metrics detailed in Table 1, excluding MoverScore, UniEval, and SimCLS. We averaged the scores across all PLMs, grouped by the weight and reranker. To ensure comparability between metrics with different scales, we applied MinMax normalization to the aggregated scores. Finally, we restructured the data into a long format and categorized the metrics according to our defined groups.

Results The results are presented in Table 2. The optimal combination weight (w) varies significantly by dataset and reference-free metric. On CNN/DM, the optimal weight for both FENICE and FIZZ is **0.75**, whereas on XSum, FENICE peaks at 0.25. Conversely, SimCLS performed best with a weight of 1.0 on both datasets, indicating that the reference-free score negatively impacted performance compared to using MBR alone. To maintain a unified setting, we selected the configuration that performed best across both datasets. We selected $w = 0.75$ as the default, as it is optimal for CNN/DM and competitive for XSum. Given the negative contribution of SimCLS, we exclude it from the final system comparison.

4.3 Experiment Results

The main results are shown in Tables 3 and 4 for the CNN/DM and XSum datasets, respectively. For settings with multiple models, i.e., `epsilon` and `beam-sim`, we averaged the results from all models. The `Baseline` refers to the non-reranked summary generated by each setting's decoding method.

The tables show that our methods **significantly dominate most of the factuality metrics** when compared with the baselines as shown by our methods achieving the highest scores and are often statistically significant against all three baselines. These results show that our method of using pseudo-references as an alternative signal works in improving factuality of a summary.

On the CNN/DM dataset, the `FENICE-0.75` configuration significantly improved the FIZZ (Fi) score from 39.36 (Baseline) to 52.44 in the `epsilon` setting. This trend is even more pronounced on the XSum dataset, where the same configuration raised the FIZZ (Fi) score from 16.91 to 27.79. Similarly, the configuration significantly improved most MENLI scores, except for MENLI-Entailment (EM) score in `beam-sim` setting. MENLI-Entailment (EM) has the highest improvement in both datasets, where the score increased from 4.46 to 10.44 in CNN/DM dataset and -31.15 to -20.36 in XSum dataset. Given that XSum is characterized by a high frequency of hallucinations due to its abstractiveness (Maynez et al., 2020), our findings confirm that consensus-based optimization can successfully filter out these errors.

Furthermore, this enhancement in factuality does not compromise the summary quality metrics, the scores remained competitive with, or significantly better than, the baselines across most settings. For CNN/DM, the improvements are mainly by the combination methods (FENICE-0.75 and FIZZ-0.75). For XSum, our methods' improvements are small but still significant. Unfortunately, for `beam-sim`, `Baseline` is the best for all quality met-

Setting	Reranker	Quality					Factuality						
		R1	R2	RL	BS	MS	EM	CM	SM	UE	Fe	Fi	SC
epsilon	Baseline	39.96	16.88	33.38	65.52	58.03	4.46	-6.62	1.63	81.37	95.52	39.36	99.74
	FENICE-0.0	40.52	17.46	34.06	66.02	58.19	5.96	-5.21	2.14	85.18	—	55.83	99.71
	FIZZ-0.0	40.54	17.81	34.23	65.91	58.17	5.18	-5.59	2.39	84.65	98.39	—	99.72
	FENICE-0.75	**40.60***	17.47*	34.11*	65.86*	**58.25*†‡**	**10.44*†‡**	**-3.82*†‡**	2.36*	83.34*	—	52.44*	99.70
	FIZZ-0.75	**40.67*†‡**	**17.87*†**	**34.36*†‡**	65.94*	**58.24*†‡**	**7.68*†‡**	**-4.46*†‡**	**2.44***	84.16*	**98.45*‡**	—	99.70
	MBR-1.0	40.36*	17.23*	33.87*	65.60*	58.19*	**11.29*†‡**	**-3.78*†‡**	—	81.41	97.35*	47.26*	99.69
	Oracle	*51.85*	*28.45*	*45.15*	*71.35*	*61.29*	*23.82*	*-0.47*	*35.02*	*95.20*	*99.80*	*84.22*	*99.89*
beam-sim	Baseline	36.37	17.34	31.64	64.78	56.79	12.26	-5.67	2.24	89.86	97.02	68.44	97.63
	FENICE-0.0	35.69	15.94	30.67	64.61	56.74	14.56	-5.49	-0.13	89.54	—	68.89	97.65
	FIZZ-0.0	35.60	16.27	30.78	64.37	56.65	12.80	-5.91	0.45	88.94	97.66	—	97.62
	FENICE-0.75	**36.50*†‡**	16.63*†‡	31.23*†‡	**64.97*†‡**	**56.90*†‡**	13.37*‡	**-4.54*†‡**	**6.60*†‡**	**90.14*†‡**	—	65.85	**97.71*†‡**
	FIZZ-0.75	36.14*‡	16.65*†‡	31.14*†‡	64.68*†‡	56.79*†‡	13.19*‡	**-4.99*†‡**	**4.59*†‡**	89.48‡	**97.80*‡**	—	97.65*‡
	MBR-1.0	**36.56*†‡**	16.65*†‡	31.14*†‡	**64.88*†‡**	**56.87*†‡**	12.03	**-4.44*†‡**	—	89.67‡	97.32*	60.60	**97.73*†‡**
	Oracle	*46.68*	*27.01*	*41.80*	*69.90*	*59.54*	*39.80*	*-0.47*	*23.46*	*96.12*	*99.86*	*92.69*	*98.62*
llm	Baseline	34.83	13.51	28.34	64.04	56.56	4.80	-2.12	21.70	92.58	98.43	24.98	99.91
	FENICE-0.0	35.15	13.77	28.66	64.20	56.64	5.53	-2.05	21.19	92.63	—	31.12	99.90
	FIZZ-0.0	35.26	13.80	28.74	64.23	56.66	5.13	-2.27	20.92	92.55	98.89	—	99.90
	FENICE-0.75	**35.31*†**	**14.11*†‡**	28.71*	**64.38*†‡**	56.66*†	**5.95*†‡**	**-1.60*†‡**	**31.14*†‡**	**92.91*†‡**	—	28.50*	99.91*†‡
	FIZZ-0.75	**35.36*†‡**	**13.98*†‡**	**28.80*†‡**	**64.34*†‡**	**56.68*†‡**	5.34*‡	**-1.95*‡**	**25.08*†‡**	**92.70*‡**	**98.90***	—	99.90‡
	MBR-1.0	35.15*	**14.07*†‡**	28.52*	**64.36*†‡**	56.63*	**6.02*†‡**	**-1.51*†‡**	—	**92.99*†‡**	98.50	24.73	**99.92*†‡**
	Oracle	*40.88*	*18.83*	*34.05*	*67.03*	*57.93*	*14.81*	*-0.45*	*49.41*	*95.63*	*99.81*	*55.71*	*99.95*

Table 3: Results for each metric on the CNN/DM dataset. Underline indicates the highest scores for each metric and **bold** indicates better scores than all baselines. *, †, and ‡ represent the statistical significance against Baseline, FENICE-0.0, and FIZZ-0.0, respectively (See §4.1). "—" indicates the skipped settings because the metrics used in the reranking and evaluation are identical. Each abbreviation represents the following metric, **R1** - ROUGE-1, **R2** - ROUGE-2, **RL** - ROUGE-L, **BS** - BERTScore, **MS** - MoverScore, **EM** - MENLI-Entailment, **CM** - MENLI-Contradiction, **SM** - MENLI-Summarization, **UE** - UniEval-Overall, **Fe** - FENICE, **Fi** - FIZZ, and **SC** - SimCLS.

rics. However, considering that MENLI scores (EM, CM, and SM) have improvements when evaluating against the "gold" references, the summaries chosen by our method do not diverge from them.

To establish a theoretical upper bound within the generated candidate pool, we calculated `Oracle` scores for each metric by selecting the candidate from the pool that maximizes the respective metric against the "gold" references or the source document. While the gap is small for some metrics, most show a big difference, with the `Oracle` scores often being more than double of our method's best scores. This holds true for both the CNN/DM and XSum datasets. This gap indicates that, despite the gains over the baselines, there is still substantial room for improvement in identifying the best possible candidate summary from a given pool. Our additional results (Appendix G) further reinforce the robustness of our method, demonstrating its effectiveness regardless of the sampling strategy or generation model.

Our preliminary results on weight sensitivity (see §4.2) indicate that balancing the weight w between MBR and reference-free metric is necessary to get the best overall performance on both dataset. For the CNN/DM dataset, FENICE reaches its peak performance at both $w = 0$ and $w = 0.75$, whereas FIZZ and SimCLS show a more linear improvement as w increases, peaking at 0.75 and 1, respectively. Conversely, on the XSum dataset, all three models exhibit higher sensitivity to the weight parameter, with peak performance generally occurring at higher values of w, suggesting that MBR are particularly crucial for this dataset.

In this study, we treat candidates $\mathcal{Y}$ and pseudo-references $\mathcal{R}$ as different sets. This raises a question whether the performance gains stems from quality disparities between the candidate pool and the pseudo-reference pool. Our preliminary study using multiple decoding strategies to generate the candidates against epsilon-generated pseudo-references confirms that the improvements achieved by ConSUM are not derived from utilizing superior references, but from the consensus mechanism itself. More detail about our investigation is in Appendix A.2.

Lastly, we reported the average length of the generated summaries (see Appendix E) and some sample comparison summaries (see Appendix H) to see the result qualitatively. Our findings show that PLMs and LLM have different length of summaries. LLM generated more than 100 tokens for

Setting	Reranker	Quality					Factuality						
		R1	R2	RL	BS	MS	EM	CM	SM	UE	Fe	Fi	SC
epsilon	Baseline	38.27	15.91	30.60	75.20	59.38	9.50	-31.15	-23.49	85.08	45.67	16.91	98.50
	FENICE-0.0	38.67	16.29	30.96	75.47	59.44	17.41	-24.20	-17.95	88.13	—	28.84	98.41
	FIZZ-0.0	38.12	15.85	30.56	75.14	59.28	14.50	-27.22	-20.10	87.12	64.70	—	98.39
	FENICE-0.75	**38.68***‡	**16.44***†‡	**31.18***†‡	**75.66***†‡	**59.45***‡	**27.04***†‡	**-20.36***†‡	**-16.40***†‡	88.01*‡	—	27.79*	98.32
	FIZZ-0.75	38.41‡	16.17*‡	30.91*‡	75.41*‡	59.37‡	**21.03***†‡	**-23.57***†‡	-18.07*‡	87.63*‡	**66.68***‡	—	98.34
	MBR-1.0	38.57*‡	**16.38***‡	**31.11***†‡	**75.59***†‡	59.39‡	**28.20***†‡	**-18.07***†‡	—	87.53*‡	63.02*	26.21*	98.31
	Oracle	*53.82*	*31.35*	*46.79*	*81.63*	*63.70*	*43.60*	*-5.32*	*17.78*	*94.11*	*86.48*	*51.10*	*99.21*
beam-sim	Baseline	41.78	20.21	34.64	75.88	59.87	16.62	-27.06	-18.88	86.26	54.74	25.05	97.59
	FENICE-0.0	37.96	16.21	30.04	74.25	58.78	19.84	-22.05	-16.06	87.14	—	33.23	97.60
	FIZZ-0.0	37.27	15.68	29.53	73.74	58.57	16.79	-24.40	-17.94	85.92	65.08	—	97.54
	FENICE-0.75	38.48†‡	16.41†‡	30.23†‡	74.39†‡	59.02†‡	14.67	**-19.55***†‡	**-8.17***†‡	**87.22***‡	—	27.51*	**97.69***†‡
	FIZZ-0.75	37.64‡	15.92‡	29.77‡	73.90‡	58.69‡	16.76	-22.16*‡	**-13.80***†‡	86.35‡	**65.54***‡	—	97.57‡
	MBR-1.0	38.13†‡	16.02‡	29.67	74.03‡	58.89†‡	11.86	**-18.49***†‡	—	86.17‡	55.52*	22.90	**97.72***†‡
	Oracle	*54.10*	*32.27*	*47.60*	*81.15*	*63.31*	*49.79*	*-5.26*	*13.82*	*93.72*	*87.59*	*58.15*	*98.65*
llm	Baseline	19.03	5.11	13.19	61.83	53.73	0.84	-6.02	8.72	93.08	88.04	22.11	99.89
	FENICE-0.0	19.08	5.05	13.19	61.83	53.72	0.82	-5.31	9.37	93.24	—	27.38	99.89
	FIZZ-0.0	19.07	5.03	13.19	61.84	53.72	0.78	-5.63	8.46	92.99	90.41	—	99.89
	FENICE-0.75	**19.09**	**5.16**†‡	13.13	**61.91***†‡	**53.74**†‡	**1.42***†‡	**-3.38***†‡	**19.09***†‡	93.22*‡	—	24.35*	99.89*†‡
	FIZZ-0.75	**19.09**	5.06‡	13.18	**61.87**‡	53.73	**0.86**‡	**-4.44***†‡	**13.40***†‡	93.04	90.36*	—	99.89‡
	MBR-1.0	19.03	**5.17***†‡	13.05	**61.90***†‡	**53.74**	**1.72***†‡	**-2.91***†‡	—	93.14‡	88.42*	21.42	**99.90***†‡
	Oracle	*23.69*	*8.33*	*17.04*	*64.43*	*54.63*	*4.06*	*-0.82*	*32.85*	*95.79*	*96.49*	*50.80*	*99.93*

Table 4: Results for each metric on the XSum dataset. Underline indicates the highest scores for each metric and **bold** indicates better scores than all baselines. *, †, and ‡ represent the statistical significance against Baseline, FENICE-0.0, and FIZZ-0.0, respectively (See §4.1). "—" indicates the skipped settings because the metrics used in the reranking and evaluation are identical. The abbreviations are the same as those in Table 3.

both datasets, while PLMs generated around 80 and 30 tokens for CNN/DM and XSum, respectively. With limited length, the main focus of the PLM-generated summaries changed w.r.t source. On the other hand, LLM-generated summaries are not limited by the prompt, hence the changes include additional point, and correction of factual error.

4.4 Effect of the Amount of Extracted Facts

We investigated the different behaviors between FENICE and FIZZ by analyzing the relationship between the number of Atomic Content Units (ACUs) extracted and the final factuality score. Figure 3 shows a representative plot for this analysis, using summaries from the BART model on the CNN/DM dataset, since we observed similar patterns across all settings.

Our analysis revealed that there is little to no correlation between the number of ACUs in a summary and the final factuality score. A summary with more "facts" is not necessarily rated as more factual. In addition, FENICE and FIZZ operate at vastly different granularities. FENICE typically extracts and evaluates a small number of ACUs (usually 3-6) to determine the score. In contrast, FIZZ consistently decomposes a summary into a much larger number of ACUs. This difference in granularity highlights that, although they share a conceptual foundation, their mechanisms for assessing factuality are fundamentally different.

4.5 Correlation between Rerankers and Metrics

To better understand what drives the performance improvements, we analyzed the correlation between S_{fin} scores and each metric's scores. In addition, previous studies (Müller and Sennrich, 2021; Kovacs et al., 2024) have shown that reranking resulted in bias toward the optimized metric. One example bias is summary length, thus we also included the following factors to our correlation analysis: The number of facts extracted by FENICE and FIZZ; The length of candidate summaries and source documents.

Instead of word counts, we measured the length using the models' own tokenizers, as this reflects the input the model actually receives. The full correlation matrix is shown in Figure 4. Surprisingly, the MENLI score, which we used as the MBR utility function, shows a slight positive correlation with the summary length. This suggests that MENLI, as an NLI-based metric, may have a tendency to favor slightly longer summaries. As

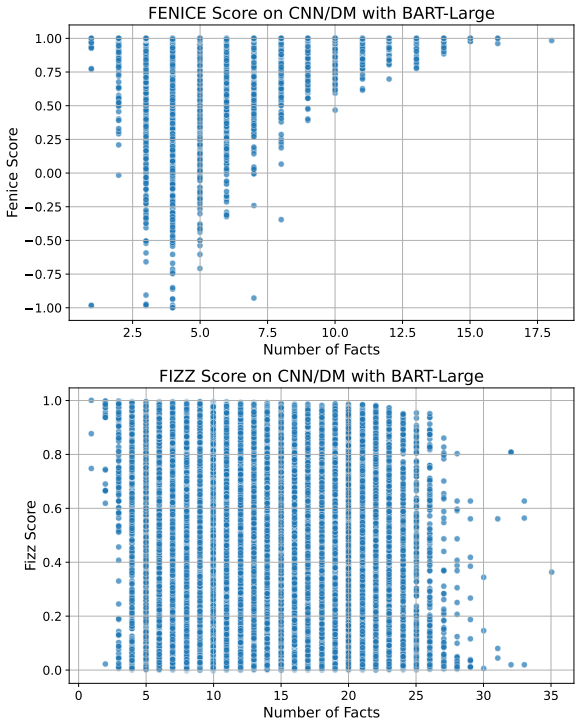

Figure 3: Correlation between the number of facts extracted using each reference-free metric and the respective factuality score for BART-generated summaries on CNN/DM. **Top:** The correlation for FENICE. **Bottom:** The correlation for FIZZ.

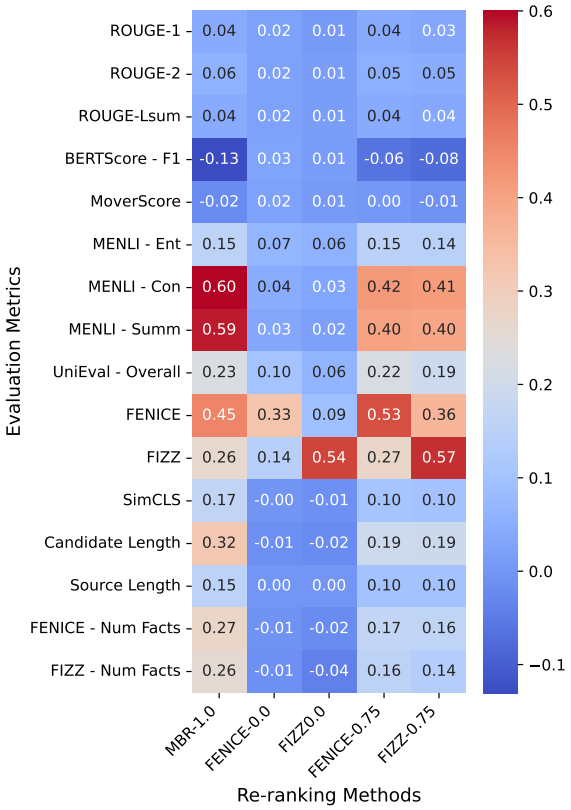

Figure 4: Correlation between rerankers and evaluation metrics. Higher numbers represent positive correlations, lower numbers represent negative correlations, and 0.0 represents no correlation.

expected, the MBR score is highly correlated with the MENLI utility function, and the FENICE and FIZZ scores are highly correlated with their own combined scores. Interestingly, even though our final score is dominated by the MBR component ($w = 0.75$), combining it with a reference-free metric increases the correlation with that metric. For example, the correlation between ConSUM (FENICE-0.75) to FENICE metric is higher compared to MBR (MBR-1.0) and FENICE reranker (FENICE-0.0) alone. This suggests that the combination helps to reduce MENLI's inherent self-correlation bias while boosting the influence of the desired reference-free metric.

5 Human Evaluation

To validate our findings beyond automatic metrics, we conducted a human evaluation to directly assess which system's outputs are preferred by people. We randomly sampled 50 source documents from the CNN/DM test set. For each source document, we presented annotators with 5 summaries to evaluate: the human-written/gold reference; the top-ranked summary from our best systems: FENICE-0.75, FIZZ-0.75, and MBR-1.0; and the top-ranked summary from the Baseline (w/o using MBR). All system-generated summaries were sourced from the BART model using the `beam-sim` setting, as this configuration yielded the best average performance in our automatic metric tests. Further discussion is presented in Appendix G.

We recruited annotators on Amazon Mechanical Turk, with 3 unique annotators evaluating each set of summaries. They were assigned two tasks: (1) Annotators rated each of the 5 summaries on a 1–5 Likert scale across three criteria: Informativeness, Factuality, and Fluency (1 = "Strongly Disagree", 5 = "Strongly Agree"). The definition of each criteria is in Appendix I. (2) Annotators ranked the same 5 summaries from best (1) to worst (5). We explicitly permitted them to assign duplicate ranks (ties) if they judged two or more summaries to be of equal quality. To ensure high-quality annotations, we recruited only MTurk workers who met the following criteria: HIT Approval Rate $\geq 90\%$ and Number of HITs Approved ≥ 50. Details is in Appendix J.

To validate our human evaluation results, we first measured the Inter-Annotator Agreement (IAA) on

System	Fact.	Info.	Fluency	Overall
Gold	4.57	3.32	3.87	3.92
MBR-1.0	4.74	**4.31**	4.66	4.57
FENICE-0.75	**4.87**	4.30	**4.73**	**4.63**
FIZZ-0.75	4.77	3.99	4.67	4.48
Baseline	4.79	4.19	4.71	4.56

Table 5: Human evaluation results for different systems based on the aspects. The overall score is calculated by averaging all aspects of each system. **Fact.** - Factuality and **Info.** - Informativeness.

the ranking task. We calculated the *Kendall's Tau* correlation (Kendall, 1948) between every pair of annotators for each sample, took the maximum of these pairwise scores, and then averaged the results across all 50 samples. This yielded an average correlation score of 0.703, indicating a strong level of agreement among the annotators and confirming the reliability of our findings.

The results for the aspect-based ratings (Informativeness, Factuality, and Fluency) are shown in Table 5. The FENICE-0.75 system achieved the highest overall average scores, making it the most preferred system when considering all aspects together. The aspect-level scores reveal each system's strengths and weaknesses. The Baseline system, while very competitive, also showed a slight weakness in informativeness compared to our method. Interestingly, the gold reference was rated poorly on both informativeness and fluency. We believe that this aligns with existing result (Zhang et al., 2024) that denote the low quality of gold references for CNN/DM.

The ranking results, visualized in Figure 5, provide a direct comparison of which summaries annotators preferred overall. FENICE-0.75 and MBR-1.0 emerge as the clear winners by having the lowest average rank (lower is better) and a low standard deviation, showing strong consensus between the annotators. This shows that annotators not only ranked these systems' summaries higher but were also more consistent in their judgment with each other. The Baseline has similar average rank with FIZZ-0.75, however with bigger standard deviation, showing that the annotators have different ranking for Baseline-generated summaries. Consistent with the aspect-based ratings, the gold reference was ranked as the worst overall.

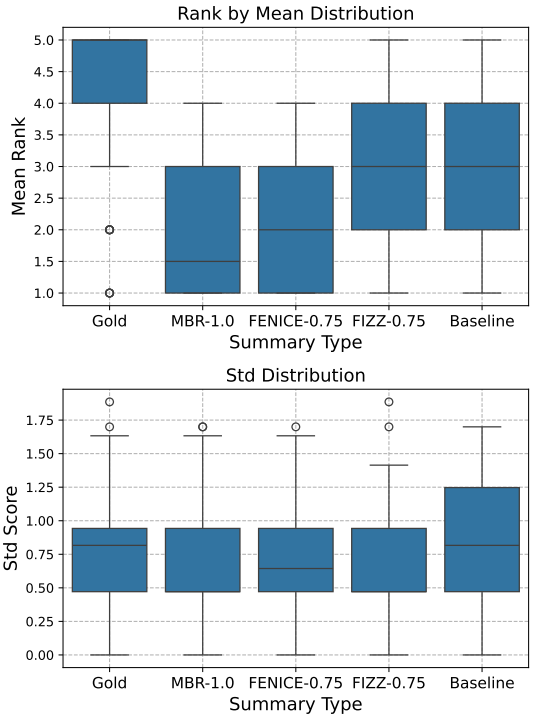

Figure 5: Rank distribution of each system. **Top:** The rank distribution based on the average ranking between annotators for each system. **Bottom:** The standard deviation of each annotator for each system.

6 Conclusion

In this work, we introduced **ConSUM**, a novel reranking framework designed to address a key limitation in summary evaluation: the reliance on either a source document or reference summaries alone. We hypothesized that the best summary is one that is not only faithful to the source but also represents the consensus of the model's own distribution. To validate this, we conducted an extensive set of experiments across multiple datasets, sampling methods, and system configurations. Our findings showed that ConSUM improves summary factuality across the board. Notably, in several settings, it achieved this without the common trade-off of sacrificing summary quality.

Finally, our human evaluation results confirmed these quantitative findings. Annotators consistently preferred the summaries produced by ConSUM over strong baselines, both in direct ranking and in aspect-based ratings. This work demonstrated that by combining source-based factuality with model-based consensus, we can generate summaries that are not only more factually reliable but also preferred by human readers.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP23H03458.

Limitations

Although our method, ConSUM, showed promising results, this work has several limitations. Our exploration of MBR settings focused only on the optimal number of candidates and pseudo-references, leaving other factors like the choice of the utility function or the pseudo-reference generation strategies unexplored. This was primarily due to MBR's $O(n^2)$ computational complexity, where increasing the number of candidates or references exponentially increases the required time and resources. Some previous studies (Trabelsi et al., 2024; Cheng and Vlachos, 2023; Natsumi et al., 2025) have proposed techniques to improve computational efficiency in MBR decoding, and these optimizations can be combined with our approach. We also encountered computational challenges with the FENICE and FIZZ metrics, which struggled with large-scale processing, hindering a more thorough parameter exploration for them. Finally, our experiments were confined to two English news datasets (CNN/DM and XSum). The inconsistent results observed between these two datasets suggest that the effectiveness of our method may vary by domain, highlighting the need for future research on a more diverse range of text types and languages.

Ethical Considerations

This study fully complies with the ACL Ethics Policy and addresses all relevant items in the Responsible Research Checklist. All resources used in this work are publicly available and properly licensed, with no concerns regarding licensing. The study does not involve or produce any harmful content. For annotation, we ensured that all rights to the artifacts were formally transferred to the authors through explicit agreements. Annotators were recruited via a crowdsourcing platform, where compensation terms were clearly stated and agreed upon in advance. All annotators were fairly compensated for their contributions. Although AI assistants were utilized for minor writing support, such as rephrasing and spell-checking, all original content was manually created by the authors. Given these points, we affirm that this work raises no ethical concerns.

References

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew Gormley. 2023. It's MBR All the Way Down: Modern Generation Techniques Through the Lens of Minimum Bayes Risk. In *Proceedings of the Big Picture Workshop*, pages 108–122, Singapore. Association for Computational Linguistics.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust Evaluation Metrics from Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Julius Cheng and Andreas Vlachos. 2023. Faster Minimum Bayes Risk Decoding with Confidence-based Pruning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12473–12480, Singapore. Association for Computational Linguistics.

Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. 2024. Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 1–11, Mexico City, Mexico. Association for Computational Linguistics.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. Mbrs: A Library for Minimum Bayes Risk Decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–362, Miami, Florida, USA. Association for Computational Linguistics.

Tanay Dixit, Fei Wang, and Muhao Chen. 2023. Improving Factuality of Abstractive Summarization without Sacrificing Summary Quality. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 902–913, Toronto, Canada. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on 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.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon Sampling Rocks: Investigating Sampling Strategies for Minimum Bayes Risk Decoding for Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 9198–9209, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hidetaka Kamigaito, Hiroyuki Deguchi, Yusuke Sakai, Katsuhiko Hayashi, and Taro Watanabe. 2025. Diversity explains inference scaling laws: Through a case study of minimum Bayes risk decoding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29060–29094, Vienna, Austria. Association for Computational Linguistics.

M.G. Kendall. 1948. *Rank Correlation Methods*. C. Griffin.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. Mitigating Metric Bias in Minimum Bayes Risk Decoding. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1063–1094, Miami, Florida, USA. 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.

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023. On Improving Summarization Factual Consistency from Natural Language Feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization. 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 1065–1072, 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.

Mathias Müller and Rico Sennrich. 2021. Understanding the Properties of Minimum Bayes Risk Decoding 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 259–272, 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.

Koki Natsumi, Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. Agreement-constrained probabilistic minimum Bayes risk decoding. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 484–493, Mumbai, India. The Asian Federation of Natural Language Processing and The 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.

Atsumoto Ohashi, Ukyo Honda, Tetsuro Morimura, and Yuu Jinnai. 2024. On the True Distribution Approximation of Minimum Bayes-Risk Decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 459–468, Mexico City, Mexico. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Preprint*, arXiv:1910.10683.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. SummaReranker: A Multi-Task Mixture-of-Experts Re-ranking Framework for Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland. Association for Computational Linguistics.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2023. Unsupervised summarization re-ranking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8341–8376, Toronto, Canada. Association for Computational Linguistics.

Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Lee, and Jungseul Ok. 2024. Multi-Dimensional Optimization for Text Summarization via Reinforcement Learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5858–5871, Bangkok, Thailand. Association for Computational Linguistics.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. FENICE: Factuality Evaluation of summarization based on Natural language Inference and Claim Extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.

Jeewoo Sul and Yong Suk Choi. 2023. Balancing Lexical and Semantic Quality in Abstractive Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 637–647, Toronto, Canada. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2023. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4265–4293, Toronto, Canada. Association for Computational Linguistics.

Firas Trabelsi, David Vilar, Mara Finkelstein, and Markus Freitag. 2024. Efficient Minimum Bayes Risk Decoding using Low-Rank Matrix Completion Algorithms. *Preprint*, arXiv:2406.02832.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *arXiv preprint*. ArXiv:1610.02424 [cs].

Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. 2024. FIZZ: Factual Inconsistency Detection by Zoom-in Summary and Zoom-out Document. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Miami, Florida, USA. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *Preprint*, arXiv:1912.08777.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Preliminary

In contrast to NMT, the application and impact of MBR decoding in the text summarization task remain relatively unexplored. This research gap

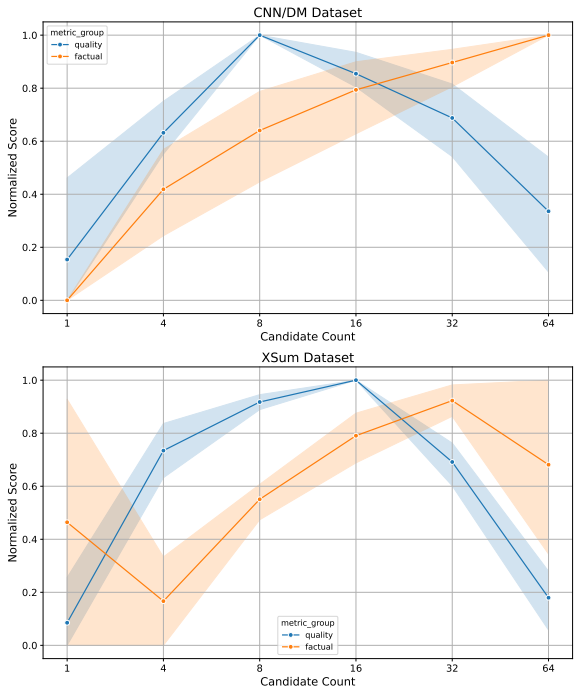

Figure 6: Average of normalized evaluation metrics based on the candidate summaries count on each dataset. **Top:** The average result on CNN/DM dataset. **Bottom:** The average result on XSum dataset. The results show the tendency of quality metrics are the same but different for factual metrics.

necessitates a preliminary study to identify the optimal hyperparameter configuration for our ConSUM method. Specifically, our study aims to determine the following: (1) the optimal number of candidates; (2) the optimal number of pseudo-references; (3) the optimal weight combination scores.

A.1 Optimal Number of Candidates

Experiment Settings Motivated by NMT findings linking performance to hypothesis count (Kovacs et al., 2024), we investigate the effect of candidate set size ($|\mathcal{Y}|$) in MBR decoding. We generated an initial count of 64 samples per document using epsilon sampling ($\epsilon = 0.02$) (Freitag et al., 2023) across three PLMs: BART, PEGASUS, and T5-Large. From this pool, we evaluated subsets of sizes 1, 4, 8, 16, 32, 64, setting the candidate set as the pseudo-reference set ($\mathcal{Y} = \mathcal{R}$) in each configuration.

Evaluation We evaluated the top-ranked summary from MBR decoding using the two groups of metrics detailed in Table 1, excluding MoverScore, UniEval, and SimCLS. We averaged the scores across all PLMs, grouped by candidate count. To ensure comparability between metrics with different scales, we applied Min-Max normalization to the aggregated scores. Finally, we restructured the data into a long format and categorized the metrics according to our defined groups.

Results The results are presented in Figure 6. In terms of quality metrics, both datasets exhibit a similar trend: performance peaks at 8 and 16 candidates for CNN/DM and XSum, respectively, before gradually declining. However, trends in factuality metrics diverge. On CNN/DM, factuality scores generally improve as the candidate count grows. Conversely, on XSum, scores rise between 4 and 32 candidates but drop thereafter. This disparity likely stems from dataset characteristics: CNN/DM's extractive nature benefits from larger consensus pools, whereas XSum's highly abstractive summaries may accumulate noise with excessive candidates. Balancing these optima, we select **16 candidates** as the default for all subsequent experiments.

A.2 Optimal Number of Pseudo-References

Experiment Settings Fixing the candidate count at the optimal $|\mathcal{Y}| = 16$ (Appendix A.1), we investigated the impact of the pseudo-reference set size ($|\mathcal{R}|$). We generated 16 candidate summaries using three strategies: Epsilon Sampling ($\epsilon = 0.02$) and Diverse Beam Search (DBS) for PLMs, and Nucleus Sampling for Llama-3. For DBS, we evaluated configurations from prior work (Dixit et al., 2023; Liu and Liu, 2021) alongside our custom settings (detailed in Table 6). We compared two reference scenarios: using the candidate set itself as the reference ($\mathcal{Y} = \mathcal{R}$), and utilizing the larger pre-generated pools from Appendix A.1 as an external reference set.

Evaluation We employed an evaluation similar to that in Appendix A.1. For PLMs, we averaged scores across all models, grouped by the type of pseudo-references used. We then applied the same Min-Max normalization and categorized the results according to our defined metric groups.

Results The results, presented in Figure 7, show varying outcomes depending on the sampling strategy. However, a clear pattern emerges for factuality metrics across all settings and datasets: **performance improves as the number of pseudo-references increases**. This aligns with the principle that a larger evidence set provides a more stable Monte Carlo approximation of the model's true

Name	Sampling Used	Models	Hyperparameter	Value Changed	Note
epsilon	Epsilon	BART PEGASUS T5-Large	epsilon cutoff do sample num beams num return sequences	0.02 True 1 16	–
beam-div	DBS	BART PEGASUS T5-Large	do sample num beams num beam groups num return sequences diversity penalty	False 4, PEGASUS=8 1 num beams 0.0	(Dixit et al., 2023)
beam-dbl	DBS	BART PEGASUS T5-Large	do sample num beams num beam groups num return sequences diversity penalty	False 8, PEGASUS=16 2 num beams 0.5	–
beam-sim	DBS	BART PEGASUS T5-Large	do sample num beams num beam groups num return sequences diversity penalty	False 16 16 num beams 1.0	(Liu and Liu, 2021)
llm	Nucleus	Llama-3	–	–	–

Table 6: Parameters and models used for each sampling setting.

posterior (Kamigaito et al., 2025). Furthermore, using epsilon-generated sets as pseudo-references is almost always superior to using the candidate set itself. This confirms that the two sets serve distinct roles: the pseudo-reference set serves as a proxy to ground truth and must be **diverse and unbiased** to accurately map the model's distribution (Kamigaito et al., 2025). In contrast, the candidate set serves as the "contestant" pool and benefits from containing multiple high-quality options, even if they are biased toward high-likelihood regions. Decoupling these sets prevents the model from reinforcing its own biases. While the trend for quality metrics is less consistent, varying by strategy and dataset (e.g., contrasting trends between CNN/DM and XSum), the signal for factuality is robust. Given this clarity, we prioritize the configuration that maximizes factuality. Therefore, a set of **64 pseudo-references generated via Epsilon Sampling** will be used as the default setting in all subsequent experiments.

A.3 Validation Results

We conclude our preliminary study by testing the optimal hyperparameters using the validation subset. Specifically, we use 16 candidates for each decoding strategy, 64 pseudo-references generated via Epsilon Sampling, and $w = 0.75$ for the combination weight.

The detailed results for all metrics are available in Table 7. For CNN/DM dataset, ConSUM improves summary quality for nearly all sampling methods, with the notable exception of Epsilon Sampling. This is likely because the epsilon candidates were being compared against a pseudo-reference set that was too similar to themselves. Our system consistently outperforms the baselines on most factuality metrics, demonstrating its effectiveness at improving factual consistency for this dataset.

For XSum, the baseline methods often achieve higher quality scores. The main exception is the LLM sampling setting, where ConSUM provides a clear benefit. This suggests that for LLM candidates, which are optimized only for source alignment, our method successfully reranks for better alignment with the model's overall distribution. Similar to the CNN/DM results, **ConSUM excels in factuality**, winning on nearly all metrics across the different sampling methods. The few exceptions likely occur when baseline summaries happen to have a higher factual overlap with the specific gold references in the validation set.

B Evaluation Metrics

We divide the evaluation metrics into two groups, Quality and Factuality as shown in Table 1. Below are the explanation for each metric.

ROUGE ROUGE (Lin, 2004) is an n-gram based metric that is widely used in summarization evalu-

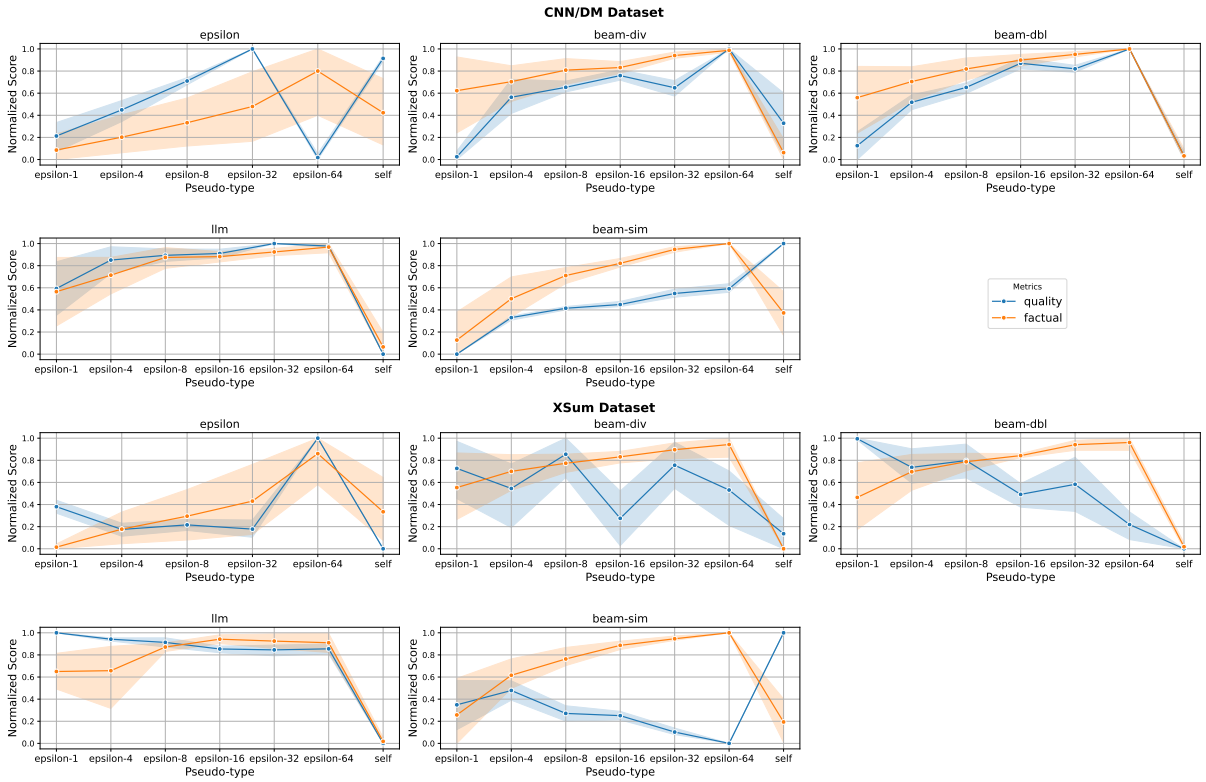

Figure 7: Average of normalized metrics based on the pseudo-reference type. Each chart represent the sampling setting for each dataset. **Top:** The average metrics for CNN/DM dataset where the sampling settings from upper left to bottom right are epsilon, beam-div, beam-dbl, llm, and beam-sim, respectively. **Bottom:** The average metrics for XSum dataset where the sampling settings from upper left to bottom right are epsilon, beam-div, beam-dbl, llm, and beam-sim, respectively.

ation. We implement the ROUGE using HuggingFace `evaluate` library.

BERTScore BERTScore (Zhang* et al., 2020) is a semantic-based similarity metric. It is also often used as a summary quality metric along ROUGE. We implement the BERTScore using HuggingFace `evaluate` library. In addition, we used the `DeBERTa-XLarge-MNLI` model (He et al., 2021), as it is the top-performing model at the time of this research.[4]

MoverScore MoverScore (Zhao et al., 2019) is a semantic-based similarity metric. Differs from BERTScore, it uses Earth Mover's Distance (EMD) on top of contextualized word embeddings.

MENLI MENLI (Chen and Eger, 2023) stands for **ME**trics from **NLI**. It is a NLI-based metric that measures a hypothesis, given the premise in three classes: Entailment – a hypothesis is true given the premise; Contradiction – a hypothesis is false given the premise; Neutral – the relationship is neither entailment nor contradiction. It can be used for various text generation tasks. We utilize the three scores provided by MENLI: entailment, contradiction, and summarization. All associated parameters are adopted from the original work. Details of each parameter are in Appendix F.

UniEval UniEval (Zhong et al., 2022) is a unified multi-dimensional evaluator. It measures all dimensions through Boolean Question Answering (QA) problem. It includes 4 different aspects and the total score: Coherence, Consistency, Fluency, Relevance, and Overall.

FENICE FENICE (Scirè et al., 2024) stands for **F**actuality **E**valuation of summarization based on **N**atural language **I**nference and **C**laim **E**xtraction. It is a two-step factuality metric comprises claim extraction and NLI alignment. It is based on the concept of Atomic Content Unit (ACU) proposed by Nenkova and Passonneau (2004). For FENICE (Scirè et al., 2024), on the claim extraction step, we diverge from the original paper's use of ChatGPT

[4]Based on the GitHub `https://github.com/Tiiiger/bert_score`, accessed on July 19, 2025.

		CNNDM									XSUM								
Setting	**Reranker**	**R1**	**R2**	**RL**	**BS**	**EM**	**CM**	**SM**	**Fe**	**Fi**	**R1**	**R2**	**RL**	**BS**	**EM**	**CM**	**SM**	**Fe**	**Fi**
epsilon	Baseline	40.48	17.29	33.99	65.60	4.95	-6.45	1.49	95.42	39.87	38.46	16.13	30.83	75.24	9.35	-30.37	-22.85	45.64	16.79
	FENICE-0.0	41.01	17.85	34.64	66.07	6.94	-4.89	2.45	—	56.25	38.80	16.42	31.14	75.52	16.80	-23.97	-17.63	—	28.70
	FIZZ-0.0	41.09	18.33	34.92	66.01	5.96	-5.29	2.60	98.33	—	38.33	15.97	30.77	75.22	14.54	-26.57	-19.44	64.82	—
	FENICE-0.75	41.06*	17.87*	34.71*	65.90*	**11.62***†‡	**-3.59***†‡	2.25*	—	53.09*	**38.93***	**16.59***	**31.39***	**75.73***	**26.61***†‡	**-19.54***†‡	**-16.09***	—	27.63*
	FIZZ-0.75	**41.18***†‡	**18.36***†	**35.00***†‡	66.04*	**8.62***†‡	**-4.16***†‡	2.59*	**98.44***‡	—	38.60*†‡	16.27*†	31.09*†‡	75.45*	**20.87***†‡	**-23.04***†‡	-17.72*	**66.63***‡	—
	MBR-1.0	40.77*	17.60*	34.42*	65.63	**12.41***†‡	**-3.59***†‡	—	97.33*	47.95*	**38.86***	**16.60***	**31.39***	**75.68**	**28.00***†‡	**-17.49***†‡	—	63.51*	26.21*
	Oracle	52.34	28.95	45.79	71.41	25.84	-0.41	34.08	99.77	84.33	53.89	31.44	46.93	81.65	43.77	-4.93	17.79	86.80	51.34
beam-div	Baseline	36.87	18.12	32.38	63.31	13.16	-5.33	2.40	97.23	72.59	42.21	20.97	35.25	73.05	17.95	-26.11	-17.85	56.17	25.55
	FENICE-0.0	36.95	18.05	32.44	63.43	13.96	-5.22	1.72	—	73.62	41.94	20.50	34.87	72.97	18.74	-25.26	-17.57	—	27.70
	FIZZ-0.0	36.87	18.05	32.39	63.32	13.49	-5.29	2.01	97.61	—	41.86	20.49	34.84	72.85	18.20	-25.77	-17.80	59.04	—
	FENICE-0.75	**37.03***†‡	18.11†‡	**32.50***†‡	**63.51***†‡	13.91*‡	**-5.04***†‡	**2.63***†‡	—	73.36*	42.04*†‡	20.53†‡	34.90*†‡	73.02*†‡	17.78*‡	**-24.71***†‡	**-15.99***†‡	—	26.94*
	FIZZ-0.75	**36.98***‡	18.10‡	**32.45***‡	63.41*‡	13.43*	**-5.03***†‡	**3.26***†‡	**97.64***	—	41.82*‡	20.42‡	34.76*‡	72.84*‡	17.46*	-25.28*†‡	**-16.57***†‡	**59.12***	—
	MBR-1.0	**37.14***†‡	**18.16**†‡	**32.51***†‡	**63.55***†‡	13.28	**-4.75***†‡	—	97.51*	71.21	42.04*†‡	20.46†‡	34.83*†‡	72.93*†‡	15.08	**-24.13***†‡	—	56.61*	24.75
	Oracle	40.80	21.54	36.22	65.52	21.40	-2.74	8.68	99.37	82.81	47.25	25.47	39.95	75.29	28.08	-17.74	-6.98	71.75	38.01
beam-dbl	Baseline	36.73	18.00	32.22	63.77	13.15	-5.29	2.80	97.28	73.06	42.14	20.95	35.21	75.97	17.91	-26.04	-17.79	56.38	25.73
	FENICE-0.0	36.56	17.48	31.93	63.77	14.31	-5.15	1.50	—	73.41	40.89	19.26	33.59	75.47	19.72	-23.42	-16.27	—	30.20
	FIZZ-0.0	36.46	17.57	31.91	63.59	13.30	-5.34	1.86	97.68	—	40.66	19.18	33.45	75.25	18.25	-24.73	-17.02	62.40	—
	FENICE-0.75	**36.83***†‡	17.70†‡	32.11†‡	**63.93***†‡	14.01*‡	**-4.66***†‡	**4.21***†‡	—	72.46	41.00*†‡	19.28†‡	33.58†‡	75.47*†‡	17.14*‡	**-22.39***†‡	**-12.61***†‡	—	27.87
	FIZZ-0.75	36.69†‡	17.70†‡	32.03†‡	63.75‡	13.45*‡	**-4.83***†‡	**4.30***†‡	**97.82***‡	—	40.62†‡	19.05†‡	33.32†‡	75.23‡	17.49*‡	-23.73*†‡	**-14.93***†‡	**62.67***	—
	MBR-1.0	**36.99***†‡	17.80†‡	32.12†‡	**63.96***†‡	13.05	**-4.43***†‡	—	97.57*	68.90	40.87*†‡	19.03†‡	33.28†‡	75.28*†‡	14.03	**-21.67***†‡	—	57.02*	24.34
	Oracle	43.28	23.89	38.63	67.41	27.73	-1.36	14.99	99.73	88.64	50.26	28.35	43.20	79.36	36.38	-11.74	1.46	80.13	47.13
beam-sim	Baseline	36.75	17.71	32.09	64.77	13.01	-5.44	2.29	96.99	68.63	41.87	20.42	34.81	75.93	16.69	-26.81	-18.56	54.51	24.79
	FENICE-0.0	35.86	16.18	30.94	64.48	15.13	-5.47	-0.24	—	69.21	38.14	16.43	30.26	74.34	19.76	-21.65	-15.67	—	32.99
	FIZZ-0.0	35.85	16.55	31.13	64.31	13.51	-5.67	0.48	97.61	—	37.45	15.92	29.74	73.82	16.46	-24.14	-17.62	65.25	—
	FENICE-0.75	**36.82**†‡	16.98†‡	31.63†‡	**64.92***†‡	14.00*‡	**-4.37***†‡	**6.35***†‡	—	66.27	38.59†‡	16.54†‡	30.31†‡	74.43*†‡	14.79*‡	**-19.08***†‡	**-7.85***†‡	—	27.32
	FIZZ-0.75	36.51†‡	17.05†‡	31.60†‡	64.67†‡	13.76*‡	**-4.77***†‡	**4.54***†‡	**97.78***‡	—	37.81†‡	16.16†‡	30.00†‡	73.98†‡	16.54*‡	-21.88*†‡	**-13.58***†‡	**65.76***	—
	MBR-1.0	**36.98***†‡	17.07†‡	31.64†‡	**64.88***†‡	12.59	**-4.34***†‡	—	97.22*	61.13	38.33*†‡	16.26†‡	29.86†‡	74.09*†‡	11.97	**-18.16***†‡	—	54.97*	22.75
	Oracle	46.86	27.29	42.09	69.82	41.07	-0.42	22.69	99.83	92.77	54.17	32.40	47.78	81.17	49.88	-4.98	14.10	87.98	58.28
llm	Baseline	35.78	13.89	29.21	64.18	5.30	-2.03	20.95	98.39	24.70	19.21	5.18	13.26	61.89	0.83	-6.11	8.68	88.27	22.32
	FENICE-0.0	36.05	14.07	29.48	64.31	6.05	-1.90	20.96	—	30.80	19.24	5.09	13.24	61.91	0.85	-5.34	8.85	—	27.76
	FIZZ-0.0	36.20	14.17	29.60	64.36	5.41	-2.05	20.73	98.88	—	19.24	5.06	13.25	61.90	0.79	-5.67	8.16	90.59	—
	FENICE-0.75	**36.22***†	**14.45***†‡	29.54*	**64.53***†‡	**6.24***†‡	**-1.50***†‡	**30.36***†‡	—	28.02*	19.20*†	**5.20***†‡	13.17*	**61.97***†‡	**1.28***†‡	**-3.50***†‡	**18.98***†‡	—	24.80*
	FIZZ-0.75	**36.27***†‡	**14.33***†‡	**29.63***†	**64.46***†‡	5.67*‡	**-1.75***†‡	**24.79***†‡	98.86*	—	19.24*†‡	5.10*†‡	13.24*†	**61.94***†‡	**0.89***‡	**-4.50***†‡	**13.17***†‡	90.56*	—
	MBR-1.0	36.12*	**14.50***†‡	29.40*	**64.51***†‡	**6.41***†‡	**-1.42***†‡	—	98.50*	24.24	19.14*	**5.21***†‡	13.11*	**61.97***†‡	**1.59***†‡	**-3.19***†‡	—	88.19*	21.68
	Oracle	41.77	19.26	34.95	67.12	15.74	-0.39	48.60	99.80	55.72	23.84	8.35	17.11	64.48	3.91	-0.92	32.57	96.45	50.74

Table 7: Results on the validation set for each metric, separated by dataset. The bold scores are the highest value scores for each metric. "—" indicates the skipped settings because the metrics used in the reranking and evaluation are identical. Each abbreviation represent each metric, **R1** - ROUGE-1, **R2** - ROUGE-2, **RL** - ROUGE-Lsum, **BS** - BERTScore, **EM** - MENLI-Entailment, **CM** - MENLI-Contradiction, **SM** - MENLI-Summarization, **Fe** - FENICE, and **Fi** - FIZZ.

and instead use the publicly available T5 distillation model provided by the authors[5].

FIZZ FIZZ (Yang et al., 2024) stands for **F**actual **I**nconsistency Detection by **Z**oom-in Summary and **Z**oom-out Document. Similar to FENICE, it is a two-step factuality metric comprises of atomic fact decomposition and NLI alignment. However, it provides more interpretability through comparison at fine-grained atomic fact level. For our FIZZ (Yang et al., 2024) setup, we use the Orca-2 model[6] as the decomposer and set the granularity level to 3G.

SimCLS SimCLS (Liu and Liu, 2021) is a contrastive learning-based scoring model based on RoBERTa, originally proposed for candidate reranking. We trained custom SimCLS models by generating candidates from the training subset of each dataset for every PLM. Specifically, we trained a distinct model for each DBS configuration, resulting in a total of 18 models (2 datasets x 3 models x 3 DBS configuration). As illustrated in Figure 8, the version trained using the **`beam-div` setting** demonstrated the most consistent performance across datasets. Therefore, we utilize this specific SimCLS model for all relevant experiments.

C Summarization Models Details

BART BART stands for Bidirectional and Auto-Regressive Transformer (Lewis et al., 2020). It is a denoising autoencoder that is trained by corrupting text with an arbitrary noising function and learning to reconstruct the original text. This objective makes the model excels in text generation tasks, especially abstractive summarization. Our study used the publicly available `BART-Large` model, fine-tuned to the respective dataset, CNN/DM[7] and XSum [8]

PEGASUS PEGASUS stands for Pre-training with Extracted Gap-sentences for Abstractive Summarization (Zhang et al., 2020). It is a transformer-based model specifically designed for abstractive

[5] https://huggingface.co/Babelscape/t5-base-summarization-claim-extractor

[6] https://huggingface.co/microsoft/Orca-2-7b

[7] https://huggingface.co/facebook/bart-large-cnn

[8] https://huggingface.co/facebook/bart-large-xsum

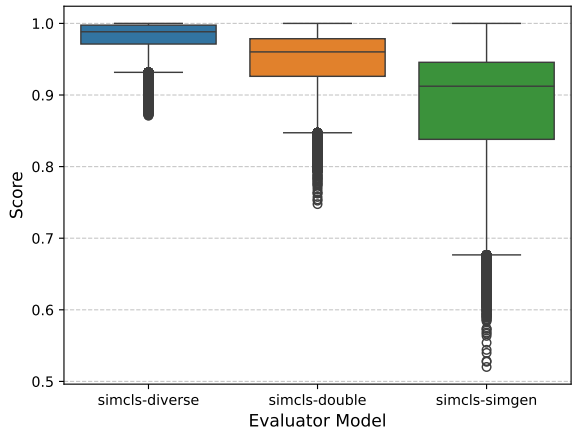

Figure 8: Comparison between SimCLS models by DBS setting

summarization through Gap Sentence Generation (GSG) objective. This objective forces the model to learn a high-level understanding of the source text. Our study used the publicly available PEGASUS model, fine-tuned to the respective dataset, CNN/DM[9] and XSum[10]

T5 T5 stands for Text-to-Text Transfer Transformer (Raffel et al., 2023). It is a model that unifies all NLP tasks into a single text-to-text format. Every task is reframed as problem of generating a target text string from an input text string with a specific prefix to allow the model to be fine-tuned to a wide variety of tasks. As there were no reliably-performing, publicly available T5 checkpoints for these summarization datasets at the time of our research, we finetuned the `t5-large` model ourselves. The process was conducted separately for each dataset, running for 5 epochs with a learning rate of 1×10^{-4} on two RTX A6000 GPUs. The performance of our fine-tuned models on the validation sets is presented in Table 8.

Llama-3 Llama-3 (Grattafiori et al., 2024) is a family of Large Language Models (LLMs) developed by Meta. It is a decoder-only transformer that is pre-trained on a massive and diverse dataset. The large amount of parameters enables the model to excel at complex reasoning and instruction following. Specifically, we used the publicly available `Llama-3-8B-Instruct` model[11]. The prompt used is in Appendix K

[9] https://huggingface.co/google/pegasus-cnn_dailymail

[10] https://huggingface.co/google/pegasus-xsum

[11] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Metric	XSUM	CNN/DM
ROUGE-1	36.47	40.73
ROUGE-2	14.29	17.58
ROUGE-L	28.69	27.27
ROUGE-Lsum	28.70	33.94

Table 8: Validation results for the T5-Large model after finetuning on the XSUM and CNN/DM datasets.

D Sampling Settings

we explore two types of decoding on the pretrained models: Diverse Beam Search (DBS) (Vijayakumar et al., 2018) and Epsilon Sampling (Hewitt et al., 2022). In addition, we explore using Large Language Model (LLM) as the summarization model for our method. For DBS, we test multiple parameter configurations inspired by previous work (Dixit et al., 2023; Liu and Liu, 2021), in addition to our own settings, which doubled the parameters' value from `beam-div` setting. For the LLM-based method, we use the Llama-3 model to generate 16 candidate summaries per source document. The detailed configurations for each sampling method are shown in Table 6.

E Generated Summary Length

The length of model-generated summaries are reported in Table 9. In this study, we purposely do not apply length restriction to emulate the average usage of summary generation using the respective models.

F MENLI Parameters

We utilize the three scores provided by MENLI: entailment, contradiction, and summarization as evaluation metrics. In addition, we use the same MENLI-summarization as the utility function for our MBR decoding method implemented using the `mbrs` (Deguchi et al., 2024) library. All associated parameters are adopted from the original work (Chen and Eger, 2023) and the parameters are in Table 10.

G Additional Test Subset Results

As shown in Appendix D, we experimented with 5 different sampling settings. We also explained the definition of each sampling setting in Appendix D. We show the performance for all 5 settings in Table 11 for the CNN/DM dataset and Table 12

Setting	Model	Dataset	
		CNN/DM	XSum
epsilon	bart	79.51	26.81
	pegasus	65.41	24.47
	t5-large	89.33	30.61
beam-sim	bart	76.43	26.62
	pegasus	73.23	23.81
	t5-large	20.89	20.71
beam-dbl	bart	78.69	25.84
	pegasus	83.17	25.14
	t5-large	20.89	20.74
beam-div	bart	78.33	25.61
	pegasus	79.64	24.17
	t5-large	20.89	20.71
llm	llama-3	138.64	115.73

Table 9: Average length of model-generated summaries across different settings and datasets.

for the XSum dataset. In addition, we included the performance of divided by setting, model, and reranker in Table 13 and Table 14 and divided by setting and model in Table 15 and Table 16 for CNN/DM and XSum dataset, respectively.

The additional results using `beam-div` and `beam-dbl` exhibit trends similar to, or better than, the `beam-sim` setting. As detailed in Tables 11 and 12, both configurations outperform the baselines, particularly on the CNN/DM dataset, where our method achieves superior ROUGE scores across all metrics. Furthermore, a model-wise analysis demonstrates consistent improvements, corroborating **the robustness of our method regardless of the underlying generation model**.

H Example Summaries

We sampled the summaries from both BART model and LlaMa-3 model to represent pre-trained model and LLM. Table 17 and Table 18 show the generated summaries using BART and `beam-sim` as the setting. Due to the fine-tuning of the model to the respective datasets, the generated summaries are limited in length. Hence, the changes in the generated summaries are mainly the main focus of the summary. On the other hand, LlaMa-3 summaries are not limited in length, thus the main difference in the generated summaries are the improvement in factual accuracy and addition of key details.

Parameter	ent	con	sum
`direction`	`rh`	`rh`	`hr`
Formula	e	$-c$	$e-c$
`nli_weight`	1.0	1.0	1.0
`combine_with`	`None`	`None`	`None`
`model`	`D`	`D`	`D`

Table 10: Parameter configurations for different MENLI variations.

I Aspect Definition

To define a clear definition of the aspects. We define them as follows:

Factuality – A summary is factual if all the information presented in it is consistent with the information in the source document. It means no details are hallucinated (made up), contradicted, or distorted from the original text. It means that the meaning from the original source is preserved.

Informativeness – Measures the inclusivity of crucial information from the document. Highly informative summary includes all key points, main ideas, and essential facts. It ensures that the reader can understand the core content without needing to refer back to the source

Fluency – A fluent summary is well-written, grammatically correct, and easy to read and understand. The text should flow naturally, use appropriate vocabulary, and have correct sentence structure, capitalization, and punctuation.

J Human Instruction

Figure 9 shows the screenshot of the human instruction from the MTurk page.

K LLM Prompt

The prompt used in LLama-3 to generate the summary is as follows. The `src` refers to the source document/the news article from the dataset:

```
{
    "role": "system",
    "content": "You are an assistant
    who replies with a summary to every
     message.",
},
{"role": "user", "content": f"Summarize
 the following text: \n\n {src}"}
```

Figure 9: MTurk Page for human evaluation.

		Quality					Factuality						
Setting	**Reranker**	**R1**	**R2**	**RL**	**BS**	**MS**	**EM**	**CM**	**SM**	**UE**	**Fe**	**Fi**	**SC**
epsilon	Baseline	39.96	16.88	33.38	65.52	58.03	4.46	-6.62	1.63	81.37	95.52	39.36	99.74
	FENICE-0.0	40.52	17.46	34.06	66.02	58.19	5.96	-5.21	2.14	85.18	—	55.83	99.71
	FIZZ-0.0	40.54	17.81	34.23	65.91	58.17	5.18	-5.59	2.39	84.65	98.39	—	99.72
	FENICE-0.75	**40.60***	17.47*	34.11*	65.86*	**58.25***†‡	**10.44***†‡	**-3.82***†‡	2.36*	83.34*	—	52.44*	99.70
	FIZZ-0.75	**40.67***†‡	**17.87***†	**34.36***†‡	65.94*	**58.24***†‡	**7.68***†‡	**-4.46***†‡	**2.44***	84.16*	**98.45***‡	—	99.70
	MBR-1.0	40.36*	17.23*	33.87*	65.60*	58.19*	**11.29***†‡	**-3.78***†‡	—	81.41	97.35*	47.26*	99.69
	Oracle	*51.85*	*28.45*	*45.15*	*71.35*	*61.29*	*23.82*	*-0.47*	*35.02*	*95.20*	*99.80*	*84.22*	*99.89*
beam-div	Baseline	36.44	17.70	31.86	64.87	56.81	12.42	-5.49	1.96	90.53	97.29	72.28	97.59
	FENICE-0.0	36.54	17.63	31.95	65.02	56.88	13.33	-5.45	1.44	90.30	—	73.52	97.61
	FIZZ-0.0	36.42	17.63	31.86	64.86	56.82	12.77	-5.42	1.61	90.10	97.62	—	97.59
	FENICE-0.75	**36.60***†‡	17.67†	**31.98***†‡	**65.05***†‡	**56.89***‡	13.21*‡	**-5.23***†‡	**2.43***†‡	90.41†‡	—	73.15*	97.61*†‡
	FIZZ-0.75	36.53*‡	17.68‡	31.92*‡	64.94*‡	56.84*‡	12.74*	**-5.22***†‡	**2.95***†‡	90.13	**97.69***‡	—	97.61*‡
	MBR-1.0	**36.74***†‡	**17.78***†‡	**32.02***†‡	**65.06***†‡	56.87*‡	12.83*	**-4.94***†‡	—	90.39†‡	**97.63***	70.88	**97.64***†‡
	Oracle	*40.42*	*21.11*	*35.73*	*66.93*	*57.83*	*20.66*	*-2.84*	*8.46*	*93.78*	*99.45*	*82.71*	*98.07*
beam-dbl	Baseline	36.25	17.55	31.66	64.76	56.74	12.32	-5.47	2.40	90.47	97.33	72.76	97.60
	FENICE-0.0	36.21	17.11	31.48	64.82	56.80	13.59	-5.35	1.26	90.01	—	73.55	97.62
	FIZZ-0.0	35.96	17.10	31.34	64.59	56.68	12.70	-5.45	1.74	89.75	97.70	—	97.61
	FENICE-0.75	**36.39***†‡	17.28†‡	31.59†‡	**64.93***†‡	**56.82***†‡	13.29*‡	**-4.88***†‡	**4.06***†‡	90.31†‡	—	72.48	**97.65***†‡
	FIZZ-0.75	36.19‡	17.23†‡	31.47‡	64.73‡	56.72‡	12.77*	**-5.05***†‡	**4.14***†‡	89.93‡	**97.81***‡	—	**97.63***‡
	MBR-1.0	**36.56***†‡	17.38†‡	31.59†‡	**64.95***†‡	56.79*‡	12.53	**-4.69***†‡	—	90.26†‡	**97.72***	68.73	**97.68***†‡
	Oracle	*42.88*	*23.41*	*38.11*	*68.13*	*58.47*	*26.56*	*-1.43*	*14.97*	*95.11*	*99.74*	*88.56*	*98.32*
beam-sim	Baseline	36.37	17.34	31.64	64.78	56.79	12.26	-5.67	2.24	89.86	97.02	68.44	97.63
	FENICE-0.0	35.69	15.94	30.67	64.61	56.74	14.56	-5.49	-0.13	89.54	—	68.89	97.65
	FIZZ-0.0	35.60	16.27	30.78	64.37	56.65	12.80	-5.91	0.45	88.94	97.66	—	97.62
	FENICE-0.75	**36.50***†‡	16.63†‡	31.23†‡	**64.97***†‡	**56.90***†‡	13.37*‡	**-4.54***†‡	**6.60***†‡	**90.14***†‡	—	65.85	**97.71***†‡
	FIZZ-0.75	36.14†‡	16.65†‡	31.14†‡	64.68†‡	56.79†‡	13.19*‡	**-4.99***†‡	**4.59***†‡	89.48‡	**97.80***‡	—	97.65*‡
	MBR-1.0	**36.56***†‡	16.65†‡	31.14†‡	**64.88***†‡	**56.87***†‡	12.03	**-4.44***†‡	—	89.67‡	97.32*	60.60	**97.73***†‡
	Oracle	*46.68*	*27.01*	*41.80*	*69.90*	*59.54*	*39.80*	*-0.47*	*23.46*	*96.12*	*99.86*	*92.69*	*98.62*
llm	Baseline	34.83	13.51	28.34	64.04	56.56	4.80	-2.12	21.70	92.58	98.43	24.98	99.91
	FENICE-0.0	35.15	13.77	28.66	64.20	56.64	5.53	-2.05	21.19	92.63	—	31.12	99.90
	FIZZ-0.0	35.26	13.80	28.74	64.23	56.66	5.13	-2.27	20.92	92.55	98.89	—	99.90
	FENICE-0.75	**35.31***†	**14.11***†‡	28.71*	**64.38***†‡	56.66*†	**5.95***†‡	**-1.60***†‡	**31.14***†‡	**92.91***†‡	—	28.50*	99.91*†‡
	FIZZ-0.75	**35.36***†‡	**13.98***†‡	**28.80***†‡	**64.34***†‡	**56.68***†‡	5.34*‡	**-1.95***‡	**25.08***†‡	**92.70***‡	**98.90***	—	99.90‡
	MBR-1.0	35.15*	**14.07***†‡	28.52*	**64.36***†‡	56.63*	**6.02***†‡	**-1.51***†‡	—	**92.99***†‡	98.50	24.73	**99.92***†‡
	Oracle	*40.88*	*18.83*	*34.05*	*67.03*	*57.93*	*14.81*	*-0.45*	*49.41*	*95.63*	*99.81*	*55.71*	*99.95*

Table 11: Results on the test set for each metric on the CNN/DM dataset. Underline indicates the highest scores for each metric and **bold** indicates better scores than all baselines. *, †, and ‡ represent the statistical significance against Baseline, FENICE-0.0, and FIZZ-0.0, respectively (See §4.1). "—" indicates the skipped settings because the metrics used in the reranking and evaluation are identical. Each abbreviation represents the following metric, **R1** - ROUGE-1, **R2** - ROUGE-2, **RL** - ROUGE-L, **BS** - BERTScore, **MS** - MoverScore, **EM** - MENLI-Entailment, **CM** - MENLI-Contradiction, **SM** - MENLI-Summarization, **UE** - UniEval-Overall, **Fe** - FENICE, **Fi** - FIZZ, and **SC** - SimCLS.

Setting	Reranker	Quality					Factuality						
		R1	R2	RL	BS	MS	EM	CM	SM	UE	Fe	Fi	SC
epsilon	Baseline	38.27	15.91	30.60	75.20	59.38	9.50	-31.15	-23.49	85.08	45.67	16.91	98.50
	FENICE-0.0	38.67	16.29	30.96	75.47	59.44	17.41	-24.20	-17.95	88.13	—	28.84	98.41
	FIZZ-0.0	38.12	15.85	30.56	75.14	59.28	14.50	-27.22	-20.10	87.12	64.70	—	98.39
	FENICE-0.75	**38.68***‡	**16.44***†‡	**31.18***†‡	**75.66***†‡	**59.45***‡	**27.04***†‡	**-20.36***†‡	**-16.40***†‡	88.01*‡	—	27.79*	98.32
	FIZZ-0.75	38.41‡	16.17*‡	30.91*‡	75.41*‡	59.37‡	**21.03***†‡	**-23.57***†‡	-18.07*‡	87.63*‡	**66.68***‡	—	98.34
	MBR-1.0	38.57*‡	**16.38***‡	**31.11***†‡	**75.59***†‡	59.39‡	**28.20***†‡	**-18.07***†‡	—	87.53*‡	63.02*	26.21*	98.31
	Oracle	*53.82*	*31.35*	*46.79*	*81.63*	*63.70*	*43.60*	*-5.32*	*17.78*	*94.11*	*86.48*	*51.10*	*99.21*
beam-div	Baseline	42.17	20.83	35.17	75.99	59.88	17.95	-26.22	-18.19	86.49	55.75	25.78	97.56
	FENICE-0.0	41.84	20.36	34.75	75.90	59.79	19.16	-25.27	-17.49	86.88	—	27.99	97.56
	FIZZ-0.0	41.74	20.28	34.72	75.81	59.75	18.44	-25.74	-17.88	86.54	58.59	—	97.55
	FENICE-0.75	41.95†‡	20.37‡	34.80‡	75.91‡	59.84†‡	18.06	**-24.71***†‡	**-15.97***†‡	**86.91***‡	—	27.21*	**97.57***†‡
	FIZZ-0.75	41.75	20.20	34.67	75.80	59.77‡	17.92	**-25.12***‡	**-16.69***†‡	86.61*‡	**58.76***	—	97.56‡
	MBR-1.0	41.98†‡	20.32	34.74	75.86‡	59.88†‡	15.33	**-24.14***†‡	—	86.73*‡	56.11	24.75	**97.59***†‡
	Oracle	*47.14*	*25.25*	*39.83*	*78.01*	*61.24*	*28.41*	*-17.80*	*-7.09*	*90.15*	*70.70*	*38.00*	*98.05*
beam-dbl	Baseline	42.09	20.80	35.12	75.91	59.86	17.94	-26.11	-18.02	86.40	55.98	25.95	97.56
	FENICE-0.0	40.70	19.15	33.39	75.39	59.45	19.93	-23.63	-16.66	87.18	—	30.45	97.56
	FIZZ-0.0	40.53	19.00	33.29	75.19	59.36	18.45	-24.89	-17.38	86.50	61.79	—	97.55
	FENICE-0.75	40.92†‡	19.16‡	33.40	75.42‡	59.56†‡	17.28	**-22.38***†‡	**-12.85***†‡	**87.19***‡	—	28.00*	**97.60***†‡
	FIZZ-0.75	40.53	18.91	33.20	75.18	59.39‡	17.84	-23.78*‡	**-15.12***†‡	86.64*‡	**62.11***‡	—	97.56‡
	MBR-1.0	40.84†‡	18.97	33.21	75.29‡	59.54†‡	14.20	**-21.79***†‡	—	86.73*‡	57.15*	24.52	**97.64***†‡
	Oracle	*50.11*	*28.10*	*43.04*	*79.32*	*62.06*	*36.43*	*-11.95*	*1.30*	*91.97*	*79.38*	*46.95*	*98.33*
beam-sim	Baseline	41.78	20.21	34.64	75.88	59.87	16.62	-27.06	-18.88	86.26	54.74	25.05	97.59
	FENICE-0.0	37.96	16.21	30.04	74.25	58.78	19.84	-22.05	-16.06	87.14	—	33.23	97.60
	FIZZ-0.0	37.27	15.68	29.53	73.74	58.57	16.79	-24.40	-17.94	85.92	65.08	—	97.54
	FENICE-0.75	38.48†‡	16.41†‡	30.23†‡	74.39†‡	59.02†‡	14.67	**-19.55***†‡	**-8.17***†‡	**87.22***‡	—	27.51*	**97.69***†‡
	FIZZ-0.75	37.64‡	15.92‡	29.77‡	73.90‡	58.69‡	16.76	-22.16*‡	**-13.80***†‡	86.35‡	**65.54***‡	—	97.57‡
	MBR-1.0	38.13†‡	16.02‡	29.67	74.03‡	58.89†‡	11.86	**-18.49***†‡	—	86.17‡	55.52*	22.90	**97.72***†‡
	Oracle	*54.10*	*32.27*	*47.60*	*81.15*	*63.31*	*49.79*	*-5.26*	*13.82*	*93.72*	*87.59*	*58.15*	*98.65*
llm	Baseline	19.03	5.11	13.19	61.83	53.73	0.84	-6.02	8.72	93.08	88.04	22.11	99.89
	FENICE-0.0	19.08	5.05	13.19	61.83	53.72	0.82	-5.31	9.37	93.24	—	27.38	99.89
	FIZZ-0.0	19.07	5.03	13.19	61.84	53.72	0.78	-5.63	8.46	92.99	90.41	—	99.89
	FENICE-0.75	**19.09**	**5.16**†‡	13.13	**61.91***†‡	**53.74**†‡	**1.42***†‡	**-3.38***†‡	**19.09***†‡	93.22*‡	—	24.35*	99.89*†‡
	FIZZ-0.75	**19.09**	5.06‡	13.18	**61.87**‡	53.73	**0.86**‡	**-4.44***†‡	**13.40***†‡	93.04	90.36*	—	99.89‡
	MBR-1.0	19.03	**5.17***†‡	13.05	**61.90***†‡	**53.74**	**1.72***†‡	**-2.91***†‡	—	93.14‡	88.42*	21.42	**99.90***†‡
	Oracle	*23.69*	*8.33*	*17.04*	*64.43*	*54.63*	*4.06*	*-0.82*	*32.85*	*95.79*	*96.49*	*50.80*	*99.93*

Table 12: Results on the test set for each metric on XSum dataset. Underline indicates the highest scores for each metric and **bold** indicates better scores than all baselines. *, †, and ‡ represent the statistical significance against Baseline, FENICE-0.0, and FIZZ-0.0, respectively (See §4.1). "—" indicates the skipped settings because the metrics used in the reranking and evaluation are identical. The abbreviations are the same as those in Table 11.

Setting	Model	Reranker	Quality					Factuality						
			R1	R2	RL	BS	MS	EM	CM	SM	UE	Fe	Fi	SC
epsilon	BART	Baseline	40.48	17.10	33.68	65.85	58.19	3.50	-6.20	3.21	85.49	96.18	39.74	99.84
		FENICE-0.0	41.05	17.71	34.39	66.29	58.37	4.78	-4.94	3.88	87.99	—	55.42	99.81
		FIZZ-0.0	41.07	18.06	34.59	66.15	58.32	4.02	-5.07	4.07	87.56	98.68	—	99.82
		FENICE-0.75	41.38	17.88	34.67	66.31	58.50	8.39	-3.31	4.13	87.49	—	51.78	99.80
		FIZZ-0.75	41.27	18.15	34.77	66.25	58.41	5.89	-4.02	4.02	87.65	98.72	—	99.81
		MBR-1.0	41.20	17.66	34.47	66.12	58.46	9.24	-3.35	—	86.53	97.64	46.55	99.80
		Oracle	52.49	28.94	45.59	71.74	61.58	19.70	-0.39	42.04	94.88	99.81	83.51	99.94
	PEGASUS	Baseline	38.78	16.06	32.56	65.04	57.65	5.43	-6.78	-1.15	76.12	94.82	38.64	99.67
		FENICE-0.0	39.31	16.61	33.23	65.46	57.76	7.07	-5.42	-0.31	81.18	—	55.60	99.65
		FIZZ-0.0	39.30	17.00	33.33	65.33	57.74	6.17	-6.08	-0.57	80.12	98.00	—	99.65
		FENICE-0.75	39.37	16.58	33.20	65.28	57.79	12.15	-3.90	0.30	78.41	—	52.61	99.63
		FIZZ-0.75	39.43	17.05	33.45	65.33	57.79	8.97	-4.76	-0.02	79.28	98.04	—	99.63
		MBR-1.0	39.07	16.28	32.93	65.02	57.72	12.99	-3.86	—	75.60	96.87	47.38	99.62
		Oracle	50.87	27.67	44.48	70.83	60.80	28.03	-0.50	26.06	95.96	99.79	85.93	99.86
	T5-Large	Baseline	40.62	17.47	33.90	65.69	58.26	4.46	-6.87	2.82	82.50	95.56	39.68	99.71
		FENICE-0.0	41.21	18.05	34.57	66.32	58.45	6.01	-5.28	2.83	86.37	—	56.46	99.68
		FIZZ-0.0	41.25	18.38	34.76	66.24	58.44	5.36	-5.62	3.66	86.26	98.48	—	99.69
		FENICE-0.75	41.06	17.96	34.46	65.98	58.46	10.79	-4.23	2.66	84.11	—	52.94	99.66
		FIZZ-0.75	41.32	18.40	34.86	66.23	58.50	8.20	-4.60	3.32	85.54	98.58	—	99.67
		MBR-1.0	40.80	17.74	34.21	65.66	58.39	11.64	-4.15	—	82.08	97.54	47.85	99.66
		Oracle	52.21	28.74	45.38	71.49	61.49	23.73	-0.51	36.94	94.76	99.79	83.21	99.88
beam-sim	BART	Baseline	42.28	20.13	35.95	66.53	58.43	3.45	-4.10	9.80	90.68	98.95	66.36	99.89
		FENICE-0.0	42.29	19.41	35.81	66.75	58.56	4.99	-4.29	4.61	90.54	—	64.17	99.81
		FIZZ-0.0	42.03	19.54	35.72	66.49	58.42	4.24	-4.40	5.51	90.08	99.19	—	99.83
		FENICE-0.75	43.09	20.19	36.33	67.06	58.69	4.01	-3.43	13.64	90.89	—	61.91	99.86
		FIZZ-0.75	42.62	20.05	36.10	66.79	58.55	3.95	-3.76	11.17	90.38	99.18	—	99.85
		MBR-1.0	43.36	20.44	36.44	67.15	58.70	3.38	-3.41	—	90.42	98.64	56.62	99.88
		Oracle	52.91	30.04	46.50	71.94	61.65	18.40	-0.42	37.09	95.65	99.85	89.75	99.93
	PEGASUS	Baseline	42.05	19.80	35.96	66.72	58.22	5.23	-5.69	3.48	93.10	98.28	57.32	99.64
		FENICE-0.0	40.96	18.21	34.57	66.35	57.98	7.43	-5.44	1.12	93.09	—	60.75	99.56
		FIZZ-0.0	41.10	18.66	34.90	66.29	57.96	6.22	-5.65	2.33	92.47	98.91	—	99.59
		FENICE-0.75	42.21	19.28	35.38	66.83	58.17	5.45	-4.27	9.92	93.53	—	55.83	99.67
		FIZZ-0.75	41.79	19.20	35.31	66.60	58.10	5.71	-4.78	7.09	93.00	98.87	—	99.64
		MBR-1.0	42.33	19.30	35.27	66.82	58.13	4.55	-3.96	—	93.53	98.03	48.71	99.71
		Oracle	51.98	28.92	45.65	71.58	60.96	25.59	-0.54	30.35	97.70	99.85	90.38	99.84
	T5-Large	Baseline	24.78	12.10	23.01	61.09	53.72	28.08	-7.23	-6.56	85.81	93.83	81.63	93.36
		FENICE-0.0	23.83	10.20	21.64	60.74	53.69	31.25	-6.75	-6.11	85.00	—	81.75	93.58
		FIZZ-0.0	23.67	10.62	21.72	60.32	53.57	27.94	-7.68	-6.49	84.27	94.87	—	93.43
		FENICE-0.75	24.20	10.43	21.97	61.02	53.83	30.64	-5.92	-3.76	86.01	—	79.81	93.59
		FIZZ-0.75	24.01	10.70	22.00	60.65	53.71	29.90	-6.45	-4.48	85.05	95.34	—	93.46
		MBR-1.0	23.98	10.21	21.71	60.67	53.79	28.15	-5.96	—	85.04	95.29	76.46	93.59
		Oracle	35.13	22.08	33.24	66.19	56.01	75.43	-0.46	2.94	95.00	99.89	97.94	96.09
llm	LLaMA-3	Baseline	34.83	13.51	28.34	64.04	56.56	4.80	-2.12	21.70	92.58	98.43	24.98	99.91
		FENICE-0.0	35.15	13.77	28.66	64.20	56.64	5.53	-2.05	21.19	92.63	—	31.12	99.90
		FIZZ-0.0	35.26	13.80	28.74	64.23	56.66	5.13	-2.27	20.92	92.55	98.89	—	99.90
		FENICE-0.75	35.31	14.11	28.71	64.38	56.66	5.95	-1.60	31.14	92.91	—	28.50	99.91
		FIZZ-0.75	35.36	13.98	28.80	64.34	56.68	5.34	-1.95	25.08	92.70	98.90	—	99.90
		MBR-1.0	35.15	14.07	28.52	64.36	56.63	6.02	-1.51	—	92.99	98.50	24.73	99.92
		Oracle	40.88	18.83	34.05	67.03	57.93	14.81	-0.45	49.41	95.63	99.81	55.71	99.95

Table 13: Results on the test set divided by setting, model, and reranker for each metric on CNN/DM dataset. Underline scores are the highest scores for each metric. "—" indicates the skipped settings because the metrics used in the reranking and evaluation are identical. The abbreviations are the same as those in Table 11.

Setting	Model	Reranker	Quality					Factuality						
			R1	R2	RL	BS	MS	EM	CM	SM	UE	Fe	Fi	SC
epsilon	BART	Baseline	37.98	15.54	30.31	75.30	59.34	9.64	-31.78	-24.35	84.92	44.58	16.70	98.83
		FENICE-0.0	38.60	16.10	30.89	75.66	59.42	18.43	-24.64	-18.73	88.31	—	28.45	98.75
		FIZZ-0.0	38.00	15.56	30.42	75.30	59.23	15.08	-27.75	-20.89	87.31	64.41	—	98.73
		FENICE-0.75	38.51	16.11	30.96	75.79	59.41	27.98	-20.43	-17.02	88.22	—	27.83	98.68
		FIZZ-0.75	38.27	15.87	30.72	75.55	59.33	21.57	-24.02	-18.48	87.89	66.55	—	98.70
		MBR-1.0	38.39	16.03	30.87	75.67	59.33	28.79	-18.04	—	87.71	62.40	26.30	98.67
		Oracle	53.84	31.24	46.79	81.76	63.67	45.39	-5.23	17.29	94.10	86.49	51.02	99.41
	PEGASUS	Baseline	40.38	18.00	32.79	76.38	60.03	12.80	-28.28	-19.67	86.17	45.23	16.40	98.22
		FENICE-0.0	40.63	18.23	32.96	76.56	60.07	22.30	-22.49	-15.27	88.86	—	27.79	98.15
		FIZZ-0.0	40.14	17.91	32.69	76.34	59.94	19.05	-25.30	-17.22	87.96	63.92	—	98.11
		FENICE-0.75	40.69	18.50	33.30	76.82	60.07	33.89	-18.87	-13.72	88.87	—	26.66	98.06
		FIZZ-0.75	40.55	18.30	33.13	76.63	60.03	27.26	-21.51	-15.30	88.48	65.99	—	98.07
		MBR-1.0	40.60	18.44	33.22	76.79	60.01	35.76	-16.47	—	88.52	62.84	25.22	98.04
		Oracle	56.42	34.49	49.72	82.90	64.74	52.59	-4.94	21.69	94.16	85.49	48.99	99.16
	T5-Large	Baseline	36.45	14.19	28.71	73.91	58.78	6.07	-33.41	-26.45	84.14	47.21	17.64	98.43
		FENICE-0.0	36.78	14.54	29.05	74.19	58.83	11.51	-25.48	-19.85	87.22	—	30.28	98.34
		FIZZ-0.0	36.20	14.08	28.56	73.78	58.66	9.37	-28.60	-22.19	86.09	65.77	—	98.32
		FENICE-0.75	36.84	14.72	29.28	74.38	58.87	19.27	-21.78	-18.47	86.94	—	28.90	98.22
		FIZZ-0.75	36.42	14.34	28.89	74.04	58.74	14.25	-25.20	-20.43	86.51	67.51	—	98.26
		MBR-1.0	36.72	14.66	29.23	74.30	58.83	20.04	-19.71	—	86.38	63.83	27.10	98.21
		Oracle	51.21	28.33	43.86	80.23	62.69	32.80	-5.78	14.35	94.06	87.46	53.30	99.07
beam-sim	BART	Baseline	43.15	20.78	35.41	77.40	60.54	14.36	-29.07	-20.56	88.69	55.14	22.07	98.73
		FENICE-0.0	39.40	16.83	30.87	75.71	59.40	18.16	-24.03	-17.56	89.19	—	29.87	98.74
		FIZZ-0.0	38.89	16.45	30.57	75.45	59.29	15.27	-26.75	-19.99	88.56	65.88	—	98.73
		FENICE-0.75	39.66	16.86	30.79	75.73	59.57	11.90	-21.36	-9.21	89.36	—	23.94	98.88
		FIZZ-0.75	39.30	16.69	30.82	75.61	59.41	14.88	-24.72	-15.93	88.88	66.34	—	98.76
		MBR-1.0	39.28	16.43	30.17	75.37	59.45	8.85	-20.32	—	88.37	55.04	19.34	98.94
		Oracle	54.90	32.52	48.12	82.20	63.90	45.34	-6.24	15.68	94.33	86.54	52.85	99.38
	PEGASUS	Baseline	45.30	23.38	37.88	78.48	61.32	20.29	-25.77	-16.05	88.76	54.43	21.57	98.01
		FENICE-0.0	40.86	18.54	32.36	76.47	59.91	22.79	-21.66	-14.40	89.54	—	29.17	98.13
		FIZZ-0.0	40.45	18.19	32.12	76.30	59.83	20.49	-23.57	-15.36	88.89	65.48	—	98.07
		FENICE-0.75	41.73	19.04	32.84	76.77	60.28	15.07	-19.67	-5.52	89.51	—	22.50	98.28
		FIZZ-0.75	40.87	18.48	32.36	76.46	59.98	19.85	-21.67	-11.24	89.15	65.51	—	98.10
		MBR-1.0	41.48	18.64	32.35	76.54	60.20	11.54	-18.78	—	88.73	53.34	17.97	98.33
		Oracle	57.37	35.66	50.90	83.28	64.95	53.17	-5.81	19.47	94.35	85.62	51.05	99.07
	T5-Large	Baseline	36.87	16.47	30.64	71.77	57.75	15.20	-26.34	-20.05	81.35	54.66	31.52	96.02
		FENICE-0.0	33.64	13.26	26.88	70.57	57.03	18.59	-20.46	-16.22	82.68	—	40.64	95.93
		FIZZ-0.0	32.46	12.40	25.91	69.48	56.61	14.62	-22.89	-18.47	80.32	63.89	—	95.83
		FENICE-0.75	34.06	13.33	27.05	70.68	57.21	17.03	-17.60	-9.78	82.80	—	36.10	95.91
		FIZZ-0.75	32.76	12.60	26.13	69.62	56.68	15.55	-20.08	-14.22	81.01	64.77	—	95.85
		MBR-1.0	33.64	13.00	26.50	70.17	57.04	15.18	-16.36	—	81.42	58.18	31.38	95.89
		Oracle	50.02	28.62	43.79	77.99	61.06	50.85	-3.74	6.31	92.50	90.63	70.57	97.50
llm	LLaMA-3	Baseline	19.03	5.11	13.19	61.83	53.73	0.84	-6.02	8.72	93.08	88.04	22.11	99.89
		FENICE-0.0	19.08	5.05	13.19	61.83	53.72	0.82	-5.31	9.37	93.24	—	27.38	99.89
		FIZZ-0.0	19.07	5.03	13.19	61.84	53.72	0.78	-5.63	8.46	92.99	90.41	—	99.89
		FENICE-0.75	19.09	5.16	13.13	61.91	53.74	1.42	-3.38	19.09	93.22	—	24.35	99.89
		FIZZ-0.75	19.09	5.06	13.18	61.87	53.73	0.86	-4.44	13.40	93.04	90.36	—	99.89
		MBR-1.0	19.03	5.17	13.05	61.90	53.74	1.72	-2.91	—	93.14	88.42	21.42	99.90
		Oracle	23.69	8.33	17.04	64.43	54.63	4.06	-0.82	32.85	95.79	96.49	50.80	99.93

Table 14: Results on the test set divided by setting, model, and reranker for each metric on XSum dataset. Underline scores are the highest scores for each metric. "—" indicates the skipped settings because the metrics used in the reranking and evaluation are identical. The abbreviations are the same as those in Table 11.

		Quality					Factuality						
Setting	**Model**	**R1**	**R2**	**RL**	**BS**	**MS**	**EM**	**CM**	**SM**	**UE**	**Fe**	**Fi**	**SC**
epsilon	BART	42.71	19.36	36.02	66.96	58.84	7.93	-3.90	9.31	88.23	98.64	63.13	99.83
	PEGASUS	40.88	18.18	34.74	66.04	58.18	11.54	-4.47	3.45	80.95	98.12	64.26	99.67
	T5-Large	42.64	19.53	36.02	66.80	58.86	10.03	-4.47	7.79	85.95	98.47	63.40	99.71
beam-div	BART	43.17	20.99	36.99	67.12	58.73	3.89	-3.99	8.68	91.59	99.21	72.44	99.85
	PEGASUS	42.71	20.90	36.79	67.00	58.37	6.03	-4.62	6.68	**93.80**	99.07	70.79	99.66
	T5-Large	25.41	12.62	23.65	61.63	53.87	32.07	-6.21	-5.31	87.02	96.83	87.25	93.51
beam-dbl	BART	43.44	21.13	37.12	67.22	58.77	4.33	-3.67	10.82	91.76	**99.36**	75.26	99.86
	PEGASUS	42.66	20.77	36.58	66.95	58.30	6.57	-4.31	9.34	93.61	99.16	73.52	99.69
	T5-Large	25.52	12.55	23.67	61.65	53.94	33.57	-5.87	-4.83	87.14	97.08	**88.11**	93.64
beam-sim	BART	**44.08**	**21.40**	**37.55**	**67.53**	**59.00**	6.06	-3.40	14.27	91.23	99.35	73.84	99.86
	PEGASUS	43.20	20.48	36.72	67.31	58.50	8.60	-4.33	9.58	93.77	99.07	70.26	99.66
	T5-Large	25.66	12.34	23.61	61.53	54.04	**35.91**	-5.78	-3.99	86.60	96.99	87.56	93.87
llm	LLaMA-3	35.99	14.58	29.40	64.65	56.83	6.80	**-1.71**	**29.13**	93.14	99.15	39.35	**99.91**

Table 15: Results on the test set divided by setting, model, and reranker for each metric on CNN/DM dataset. **Bold** and Underline represent the highest and the second highest scores, respectively. The abbreviations are the same as those in Table 11.

		Quality					Factuality						
Setting	**Model**	**R1**	**R2**	**RL**	**BS**	**MS**	**EM**	**CM**	**SM**	**UE**	**Fe**	**Fi**	**SC**
epsilon	BART	40.51	18.06	32.99	76.43	59.96	23.84	-21.70	-13.99	88.35	70.71	35.67	98.82
	PEGASUS	42.77	20.55	35.40	77.49	60.70	**29.09**	-19.69	-10.33	89.00	70.29	34.33	98.26
	T5-Large	38.66	16.41	31.08	74.98	59.34	16.19	-22.85	-15.85	87.33	72.05	37.30	98.41
beam-div	BART	43.90	21.57	36.24	77.75	60.71	17.27	-26.80	-17.82	89.47	61.13	27.12	98.74
	PEGASUS	**46.26**	**24.42**	**38.80**	**78.64**	**61.51**	22.73	-22.54	-11.40	89.50	63.62	28.01	98.12
	T5-Large	37.80	17.27	31.54	72.15	57.85	17.98	-23.10	-16.76	82.59	64.27	38.96	96.04
beam-dbl	BART	43.80	21.32	36.00	77.66	60.65	18.41	-24.85	-15.07	89.81	66.19	30.98	98.77
	PEGASUS	45.42	23.50	37.69	78.15	61.22	23.12	-20.59	-8.69	89.59	67.87	31.67	98.20
	T5-Large	37.52	16.94	31.15	72.05	57.80	19.36	-20.79	-14.41	83.16	69.27	44.17	96.09
beam-sim	BART	42.08	19.51	33.82	76.78	60.22	18.40	-21.78	-10.57	89.63	71.47	36.07	98.88
	PEGASUS	44.01	21.70	35.83	77.76	60.92	23.31	-19.56	-6.50	89.85	70.56	34.73	98.28
	T5-Large	36.21	15.67	29.56	71.47	57.63	21.00	-18.21	-11.41	83.15	73.10	**49.99**	96.13
llm	LLaMA-3	19.73	5.56	13.71	62.23	53.86	1.50	**-4.07**	**16.24**	**93.50**	**92.32**	35.22	**99.90**

Table 16: Results on the test set divided by setting, model, and reranker for each metric on XSum dataset. **Bold** and Underline represent the highest and the second highest scores, respectively. The abbreviations are the same as those in Table 11.

Source
(CNN)A photo of a baby boy being pulled from the rubble of the Nepal earthquake has become the defining image of a disaster that has devastated the country. His tiny face is coated in dust from the debris that crashed around him as the earth shook on Saturday, April 25, killing more than 5,000 people and injuring at least twice as many more. His name is Sonit Awal, and he's just four months old. On Saturday, Sonit was at home in Muldhoka, Bhaktapur, east of the capital Kathmandu, when the 7.8-magnitude quake sent tremors through Nepal and neighboring countries, according to Kathmandu Today. The house collapsed, burying the child, leaving his father Shyam Awal frantically searching for him amid the rubble. A desperate Awal called the Nepalese Army, and they too scrambled through the dirt and debris until midnight, before giving up hope and leaving. Sonit's father had lost all hope of finding his son alive, the paper reported, until he heard faint cries from below. On Sunday morning, the soldiers returned and at 10 a.m they lifted Sonit clear of the fallen rocks, beams, bricks and dust that had trapped him for 22 hours. His rescuers held him high in their hands to the sun, though his eyes were still wedged tightly shut. Sonit's face was exposed during the whole ordeal, though a hooded top and shawl protected his head and body during the cold night spent under rubble. The baby was taken to Bhaktapur Hospital and found to be uninjured; . His unlikely rescue provides some hope to a country that has experienced so much loss.

Before MBR Summary

Baseline Summary:
Sonit Awal was buried under rubble when his home collapsed in the Nepal earthquake. His father called the army, who spent 22 hours searching for him before giving up hope. The baby was rescued Sunday morning and found to be uninjured. The 7.8-magnitude quake killed more than 5,000 people and injured at least twice as many more.

FIZZ-0.0 Summary:
Sonit Awal, 4 months, was buried under rubble when his home collapsed. His father frantically searched for him for 22 hours before calling the army. The soldiers pulled him from the rubble on Sunday morning. He was uninjured. The 7.8-magnitude quake killed more than 5,000 people and injured at least twice as many.

FENICE-0.0 Summary:
Baby Sonit Awal was trapped under rubble for 22 hours after his house collapsed in Nepal's earthquake. His father called the army, which spent the night searching for him. The baby was rescued and taken to a hospital, uninjured. More than 5,000 people were killed in the earthquake, which struck on Saturday.

After MBR Summary

MBR-1.0 Summary:
Sonit Awal, 4 months old, was buried under rubble when Nepal's house collapsed. The baby's father frantically searched for him for 22 hours before calling the army. Sonit was rescued Sunday morning, uninjured, and taken to a hospital. The 7.8-magnitude quake killed more than 5,000 people and injured at least twice as many more.

FIZZ-0.75 Summary:
Sonit Awal, 4 months, was buried under rubble when his home collapsed. His father frantically searched for him for 22 hours before calling the army. The soldiers pulled him from the rubble on Sunday morning. He was uninjured. The 7.8-magnitude quake killed more than 5,000 people and injured at least twice as many.

FENICE-0.75 Summary:
Baby Sonit Awal was trapped under rubble for 22 hours after his house collapsed in Nepal's earthquake. His father called the army, which spent the night searching for him. The baby was rescued and taken to a hospital, uninjured. More than 5,000 people were killed in the earthquake, which struck on Saturday.

Table 17: Example CNN/DM dataset summaries generated by BART model using the beam-sim setting.

Source
The mix-up led to the body of Philip Bradburn being cremated instead of that of Conservative MEP Philip Bradbourn. A failure to provide written records created confusion between the two similar sounding surnames, the Heart of England NHS Trust said. The trust and Central England Co-operative apologised for the blunder. Former MEP Mr Bradbourn died at Good Hope Hospital, Sutton Coldfield on 20 December. Mr Bradburn died at University Hospital Birmingham four days later and his body was sent to a funeral directors run by Central England Co-operative. As Good Hope's mortuary was nearing capacity, a request was made by hospital staff to move four bodies - including Mr Bradbourn's - to the funeral company. The names of those to be moved were given over the phone but not followed up with an email listing their names and addresses, the report found. Updates on this story and more from Birmingham and the Black Country. The undertaker collected Mr Bradbourn's body, but four days later returned it to Good Hope, when his family requested to change his burial to a cremation. Meanwhile, Mr Bradburn's body was sent back to the hospital over concerns it had been at the funeral directors for some time. When it arrived at Good Hope from the funeral directors, paperwork carrying his surname was overwritten with the surname of the politician. Doctors looked at these papers and signed off the cremation for Mr Bradbourn but the body of Mr Bradburn was released. Central England Co-operative said: "Our priority has been to work closely with all concerned so that we can learn from this unfortunate incident and build further appropriate safeguards for the future." Dr Andrew Catto, of Heart of England NHS Foundation Trust said there had been a "rare and complex set of circumstances". He said: "We are very sorry that this incredibly distressing situation has happened."

Before MBR Summary

Baseline Summary:
An NHS trust has apologised after the body of a former Conservative MEP was mistakenly sent to the funeral of a man with the same surname.

FIZZ-0.0 Summary:
Staff at a Birmingham hospital made a "distressing mistake" when they signed off the cremation of a politician's body, a report has found.

FENICE-0.0 Summary:
An NHS trust and funeral company have apologised after the body of a former Conservative MEP was mistakenly sent to the funeral of a man with the same surname.

After MBR Summary

MBR-1.0 Summary:
An MP's body was sent to a funeral company for cremation when hospital staff mistook it for another man's, a report has found.

FIZZ-0.75 Summary:
An MP's body was sent to a funeral company for cremation when hospital staff mistook it for another man's, a report has found.

FENICE-0.75 Summary:
An MP's body was sent to a funeral company for cremation when hospital staff mistook it for another man's, a report has found.

Table 18: Example XSUM dataset summaries generated by BART model using the beam-sim setting.

Source
West Ham wonderkid Reece Oxford pushed himself closer to a long-awaited first-team debut after scoring for the Under 21s against Manchester United. The 16-year-old, who has been on the bench for the senior side this season, scored with a header in the Hammers 3-2 defeat against the Red Devils on Tuesday night. The goal will push Oxford's claims for a senior debut before the end of the season. Reece Oxford, pictured here earlier in the season, has moved closer to a first-team debut for West Ham . Oxford scored for West Ham's under 21 team in their 3-2 defeat to Manchester United on Tuesday night . Oxford signed his first professional contract earlier this season, snubbing interest from all of the Premier League's big guns. The central defender, who has been likened to Rio Ferdinand, has tipped to be a future England international. Oxford regularly trains with the first team and is hopeful Sam Allardyce will give him his first team debut before the end of campaign.

Before MBR Summary

Baseline Summary:
Reece Oxford, a 16-year-old defender for West Ham, scored a header for the Under 21 team in a 3-2 loss to Manchester United, bringing him closer to a potential first-team debut. Oxford has already been on the bench for the senior side this season and signed his first professional contract, turning down interest from top Premier League clubs. He regularly trains with the first team and hopes to make his debut under manager Sam Allardyce before the end of the season.

FIZZ-0.0 Summary:
Reece Oxford, a 16-year-old West Ham player, scored a header for the Under 21 team in a 3-2 loss to Manchester United, increasing his chances of making his senior debut before the end of the season. Oxford has been training with the first team and has signed his first professional contract, snubbing interest from top Premier League clubs. He is expected to be a future England international and hopes to make his debut under manager Sam Allardyce.

FENICE-0.0 Summary:
Reece Oxford, a 16-year-old West Ham defender, scored a goal for the Under 21 team in a 3-2 loss to Manchester United, bringing him closer to a potential first-team debut. Oxford has already been on the bench for the senior team this season and has been training with the first team. He signed his first professional contract earlier this season, despite interest from other Premier League teams, and is tipped to be a future England international.

After MBR Summary

MBR-1.0 Summary:
Reece Oxford, a 16-year-old West Ham defender, scored a header for the Under 21 team in a 3-2 loss to Manchester United, bringing him closer to a potential first-team debut. Oxford has already been on the senior team's bench this season and has signed his first professional contract, snubbing interest from top Premier League clubs. He has been tipped to become a future England international and hopes to make his first-team debut under manager Sam Allardyce before the end of the season.

FIZZ-0.75 Summary:
Reece Oxford, a 16-year-old West Ham player, scored a header for the Under 21 team in a 3-2 loss to Manchester United, increasing his chances of making his senior debut before the end of the season. Oxford has been training with the first team and has signed his first professional contract, snubbing interest from top Premier League clubs. He is expected to be a future England international and hopes to make his debut under manager Sam Allardyce.

FENICE-0.75 Summary:
Reece Oxford, a 16-year-old West Ham defender, scored a goal for the Under 21 team in a 3-2 loss to Manchester United, bringing him closer to a potential first-team debut. Oxford has already been on the bench for the senior team this season and has been training with the first team. He signed his first professional contract earlier this season, despite interest from other Premier League teams, and is tipped to be a future England international.

Table 19: Example CNN/DM summaries generated by Llama-3 model using the llm setting.

Source
The mix-up led to the body of Philip Bradburn being cremated instead of that of Conservative MEP Philip Bradbourn. A failure to provide written records created confusion between the two similar sounding surnames, the Heart of England NHS Trust said. The trust and Central England Co-operative apologised for the blunder. Former MEP Mr Bradbourn died at Good Hope Hospital, Sutton Coldfield on 20 December. Mr Bradburn died at University Hospital Birmingham four days later and his body was sent to a funeral directors run by Central England Co-operative. As Good Hope's mortuary was nearing capacity, a request was made by hospital staff to move four bodies - including Mr Bradbourn's - to the funeral company. The names of those to be moved were given over the phone but not followed up with an email listing their names and addresses, the report found. Updates on this story and more from Birmingham and the Black Country. The undertaker collected Mr Bradbourn's body, but four days later returned it to Good Hope, when his family requested to change his burial to a cremation. Meanwhile, Mr Bradburn's body was sent back to the hospital over concerns it had been at the funeral directors for some time. When it arrived at Good Hope from the funeral directors, paperwork carrying his surname was overwritten with the surname of the politician. Doctors looked at these papers and signed off the cremation for Mr Bradbourn but the body of Mr Bradburn was released. Central England Co-operative said: "Our priority has been to work closely with all concerned so that we can learn from this unfortunate incident and build further appropriate safeguards for the future." Dr Andrew Catto, of Heart of England NHS Foundation Trust said there had been a "rare and complex set of circumstances". He said: "We are very sorry that this incredibly distressing situation has happened."

Before MBR Summary

Baseline Summary:
The Heart of England NHS Trust and Central England Co-operative apologized for a mix-up that led to the cremation of the wrong body. The mistake occurred when the names of two men with similar surnames, Philip Bradburn and Philip Bradbourn, were not properly recorded. Bradburn's body was sent to a funeral director, but was later returned to the hospital due to concerns it had been there too long. Meanwhile, Bradbourn's body was cremated, despite his family requesting a burial. The trust and co-operative are working to learn from the incident and implement additional safeguards to prevent similar mistakes in the future.

FIZZ-0.0 Summary:
A mix-up led to the cremation of Philip Bradburn instead of Conservative MEP Philip Bradbourn due to a failure to provide written records and similar sounding surnames. The mistake occurred when four bodies, including Bradbourn's, were moved from Good Hope Hospital to a funeral directors run by Central England Co-operative. The undertaker collected Bradbourn's body, but it was later returned to the hospital when his family requested a cremation. Meanwhile, Bradburn's body was sent back to the hospital, and paperwork was overwritten with Bradbourn's surname, leading to the cremation of Bradburn's body. Both the Heart of England NHS Trust and Central England Co-operative have apologized for the blunder.

FENICE-0.0 Summary:
The Heart of England NHS Trust and Central England Co-operative apologized for a mistake that led to the cremation of Philip Bradburn's body instead of Conservative MEP Philip Bradbourn's. The mix-up occurred due to a failure to provide written records, resulting in confusion between the two similar-sounding surnames. Bradbourn's body was moved to a funeral director's and later returned to the hospital, while Bradburn's body was sent to the funeral director's and eventually cremated. The incident was described as a "rare and complex set of circumstances" and both organizations are working to learn from the mistake and implement further safeguards.

After MBR Summary

MBR-1.0 Summary:
The Heart of England NHS Trust and Central England Co-operative apologized for a mistake that led to the cremation of Philip Bradburn's body instead of Conservative MEP Philip Bradbourn's. The mix-up occurred due to a failure to provide written records, with similar sounding surnames contributing to the confusion. Bradburn died on December 24, and his body was sent to a funeral director, while Bradbourn died on December 20. The mistake was only discovered when Bradbourn's family requested a change from burial to cremation, and it was found that Bradburn's body had been cremated instead.

FIZZ-0.75 Summary:
A mix-up led to the cremation of Philip Bradburn instead of Conservative MEP Philip Bradbourn due to a failure to provide written records and similar sounding surnames. The mistake occurred when four bodies, including Bradbourn's, were moved from Good Hope Hospital to a funeral directors run by Central England Co-operative. The undertaker collected Bradbourn's body, but it was later returned to the hospital when his family requested a cremation. Meanwhile, Bradburn's body was sent back to the hospital, and paperwork was overwritten with Bradbourn's surname, leading to the cremation of Bradburn's body. Both the Heart of England NHS Trust and Central England Co-operative have apologized for the blunder.

FENICE-0.75 Summary:
The Heart of England NHS Trust and Central England Co-operative apologized for a mistake that led to the cremation of Philip Bradburn's body instead of Conservative MEP Philip Bradbourn's. The error occurred due to a failure to provide written records and confusion between the two similar-sounding surnames. Bradbourn's body was moved to a funeral director's, but his family later requested a cremation, while Bradburn's body was mistakenly released for cremation. The trust and co-operative are working to learn from the incident and implement additional safeguards to prevent similar mistakes in the future.

Table 20: Example XSUM summaries generated by Llama-3 model using the llm setting.