Toward Factual Summarization through Consensus and Consistency

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 [2]Chungnam National University [3]Institute of Science Tokyo
riza.setiawan_soetedjo.rs6@naist.ac.jp jingun.kwon@cnu.ac.kr
oku@first.iir.isct.ac.jp {sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

Abstract

Generating a reliable summary, which reflects the source document without any distortion to the original meaning, remains a challenge. Reranking has shown success in choosing an optimal summary, but current implementations only rely on source as guidance. To address this limitation, we propose ConSUM that reranks candidate summaries by considering two factors: **consistency** to the source document and **consensus** among other candidates. Consensus is established using Minimum Bayes Risk (MBR) decoding, while ensuring consistency by employing factuality-aware metrics. 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.

1 Introduction

Document Summarization is the task of summarizing a lengthy document while retaining its most important information. Thus, **a summary is reliable when it reflects the source document without any distortion that alters the original meaning**. In another word, factuality of the summary [1]. One method to improve factuality is reranking [2, 3], which involves generating multiple candidate summaries (hypotheses) and ranking them using metrics that reflect summarization quality.

We cannot use any gold references in reranking. Thus, reference-free metrics are frequently used to rerank the hypothesis by using, e.g., the original source document as their ground truth [4]. This limitation creates two weaknesses, where relying solely on the source is unreliable,

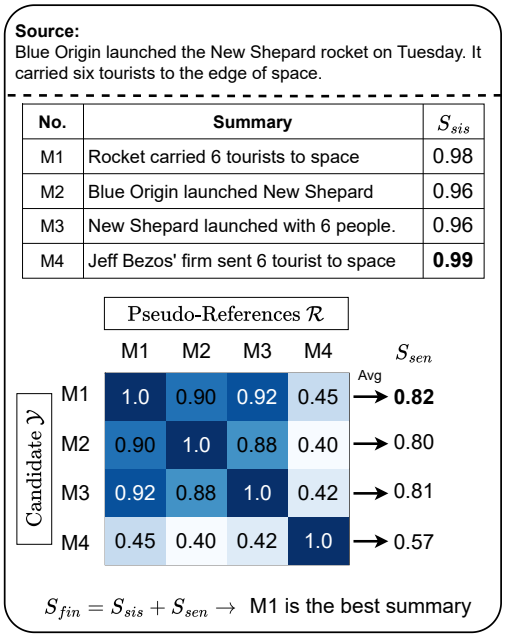

Figure 1 Case where reference-free metric is not reliable. S_{sis} represents consistency score to the source document; S_{sen} represents consensus score over among hypotheses; S_{fin} represents the final score.

as depicted in Figure 1; and because it only uses a single metric, it is prone to overfitting to the metric and inheriting its bias [5].

Inspired by Minimum Bayes Risk (MBR) decoding [6], we propose a novel reranking method to address the weaknesses, ConSUM (**Con**sistency and **Con**sensus in **Sum**marization), that combines two factors: the **consensus** among hypotheses and the **consistency** of each hypothesis to the source document for factual summary generation. Consensus is used to identify the most representative summary among the hypotheses by comparing each hypothesis to a set of pseudo-reference summaries gener-

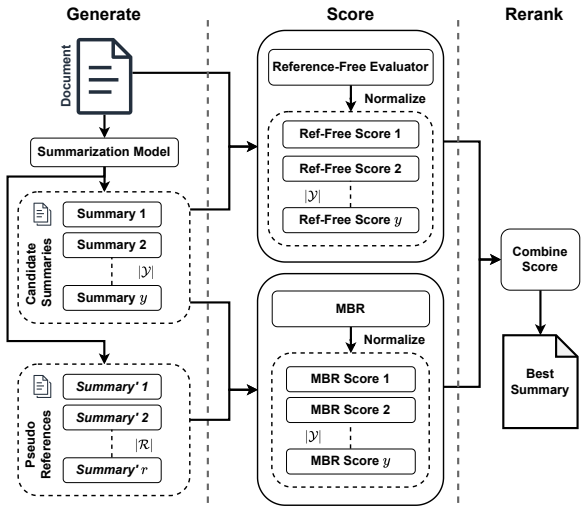

Figure 2 Overview of ConSUM comprising three steps: Generate, Score, and Rerank.

ated by the same model using a reference-based factuality metric. Consistency for the source document, meanwhile, is measured using established reference-free metrics. Finally, both scores are combined to select the summary with the highest score. Experimental results on CNN/DailyMail [7] using two types of summarization models, BART [8], and Llama-3 [9], consistently demonstrate that our method achieves superior factuality scores, showing that our system's output is preferred over baselines, supporting that combining **consensus** and **consistency** succeeded in improving the factuality of a summary.

2 Proposed Method: ConSUM

Figure 2 shows the overview of our ConSUM. It comprises three steps: (1) *Generate Hypotheses and Pseudo-references* – The summarization model generates multiple hypothesis and pseudo-references summaries for a given source document; (2) *Score Hypotheses* – Each hypothesis is scored using two distinct types of evaluators, consistency to the source document and consensus between the hypotheses and the pseudo-references; (3) *Rerank Hypotheses* – Both scores are combined using a weight, and the hypothesis with the highest score is selected as the best output.

2.1 Generate Hypotheses and Pseudo-references

This step aims to generate two pools of summaries for each source document using a summarization model, one as hypotheses and the other as pseudo-references. The hypotheses set is used to choose the best summary, and the pseudo-references set is used as a consensus reference that is further explained in §2.2.

Formally, given a source document $\boldsymbol{x}$, a decoding method δ, and a parameter θ, the summarization model p generates a set of summaries $\mathcal{S} = \{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_{|\delta|}\}$ by:

$$\boldsymbol{s}_i \sim p(\boldsymbol{s}|\boldsymbol{x}; \theta, \delta). \tag{1}$$

Given two decoding method δ_a and δ_b, the model p generates two different sets, hypotheses $\mathcal{Y}$ and pseudo-references $\mathcal{R}$. Thus:

$$\boldsymbol{s} \triangleq \begin{cases} \boldsymbol{y} \in \mathcal{Y} & \text{if } \delta = \delta_a \quad \text{(Hypothesis)} \\ \boldsymbol{r} \in \mathcal{R} & \text{if } \delta = \delta_b \quad \text{(Pseudo-reference)} \end{cases} \tag{2}$$

It is possible for $\mathcal{Y} = \mathcal{R}$ when $a = b$ [10]. However, previous work [11] has shown the importance of choosing pseudo-references that are diverse and unbiased.

2.2 Score Candidates

We utilize two distinct scores to enhance the factuality of generated summaries: consistency with the source and consensus with the pseudo-references. Consistency-based scores have been used previously [4], but relying solely on them has been shown to produce summaries that distort the original meaning of the source document [4], as depicted in Figure 1. Due to the high cost of collecting "gold" references, we leverage the pseudo-reference set $\mathcal{R}$ as a second reference signal through consensus-based scores. Specifically, $\mathcal{R}$ represents an approximation of the model's posterior distribution, allowing the decoding process to ground predictions in statistical consensus rather than relying solely on high-probability sequences. Its purpose is to mitigate any outlier information from any summary that might provide counterfactual information.

Consistency-based Scoring We define the consistency-based score that measures the factual consistency between the source document $\boldsymbol{x}$ and its hypothesis $\boldsymbol{y}_i \in \mathcal{Y}$, generated in §2.1, as follows:

$$S_{sis}(\boldsymbol{y}_i, \boldsymbol{x}) = FM(\boldsymbol{y}_i, \boldsymbol{x}), \tag{3}$$

where $FM(\boldsymbol{y}_i, \boldsymbol{x})$ is a reference-free factuality metric. In this study, we choose to use FENICE [12] and FIZZ [13] as the metric.

Consensus-based Scoring To reflect the consensus between pseudo-references in $\mathcal{R}$ on deciding the best hypothesis in $\mathcal{Y}$, we incorporate the score calculation in MBR

decoding as follows:

$$S_{sen}(\boldsymbol{y}_i, \mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum_{j=1}^{|\mathcal{R}|} u(\boldsymbol{y}_i, \boldsymbol{r}_j), \quad (4)$$

where $u(\boldsymbol{y}_i, \boldsymbol{r}_j)$ is a utility function to calculate each hypothesis $\boldsymbol{y}_i$ against a pseudo-reference $\boldsymbol{r}_j$. In this study, we choose MENLI [14] as the utility function, as it has been demonstrated to be the most effective for the summarization task.

2.3 Rerank Candidates

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

$$\begin{aligned} &S_{fin}(\boldsymbol{y}_i, \boldsymbol{x}, \mathcal{R}) \\ =&wN(S_{sen}(\boldsymbol{y}_i, \mathcal{R})) + (1-w)N(S_{sis}(\boldsymbol{y}_i, \boldsymbol{x})), \end{aligned} \quad (5)$$

where N 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:

$$\hat{\boldsymbol{y}} = \text{argmax}_{\boldsymbol{y}} S_{fin}(\boldsymbol{y}, \boldsymbol{x}, \mathcal{R}). \quad (6)$$

Previous studies about improving factuality through reranking, rely solely on reference-free metrics [4] or MBR decoding [10]. Thus, we propose combining the two approaches to obtain the most factuality-aware candidate. The best candidate is a candidate chosen by both MENLI w.r.t. pseudo-references to represent the consensus and FENICE or FIZZ w.r.t. the source document to represent the consistency.

3 Experiment Settings

We evaluate our method on a news summarization dataset, using Pre-trained Language Models (PLMs) and Large Language Models (LLMs) as the summarization models, for metrics from two aspects: Quality and Factuality.

Datasets and Models We evaluated our method on CNN/DailyMail (CNN/DM) [7], a popular news summarization dataset. It is characterized by its relatively extractive summaries, where summary sentences are often copied directly from the source article. To generate the summaries, we utilized a fine-tuned PLM, BART [8] and an LLM, Llama-3 [9]. We explored both model types because LLMs have shown weaknesses in summarization [15] that make them comparable to PLMs. See Appendix A for implementation details.

Table 1 Metrics by group and type by reference.

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

Summary Generation We generate 16 summaries per source document as the hypotheses and use Diverse Beam Search (DBS) [16] and Nucleus [17] to generate them for PLM and LLM, respectively. Inspired by this study [11], we compared two conditions, the hypotheses are used as pseudo-references ($\mathcal{Y} = \mathcal{R}$) and the pseudo-references are generated using Epsilon Sampling [18]. Specifically, 64 summaries are generated per source document as the pseudo-references using Epsilon Sampling.

Metrics The selected summaries are assessed using two metric groups: Quality and Factuality, as shown in Table 1. We used 2 reference-based metrics in the Quality group: ROUGE [19] and BERTScore [20]. In Factuality group, we utilized 3 metrics: MENLI [14] as the reference-based metric, FENICE [12], and FIZZ [13] as the reference-free metrics. Statistical significance was assessed via paired-bootstrap resampling [21] with 10,000 samples. We established a significance level of $p < 0.05$, applying the Bonferroni correction to account for multiple comparisons against the three baselines. See Appendix B for implementation details.

Hyperparameters Our preliminary experiment identified the optimal weight (w) of **0.75** (See Appendix C). This weight corresponds to a 75:25 contribution ratio of S_{sen} and S_{sis} to the S_{final} score. We denote our methods as `Scorer-`w, where w represents the weight assigned to the MBR score (S_{sen}) and `Scorer` refers to either FENICE or FIZZ. Consequently, $w = 0.0$ serves as a baseline using only the named scorer, while $w = 1.0$ uses only the MBR score and is denoted simply as `MBR`. Finally, $0 < w < 1$ indicates a linear combination between the named scorer and MBR score.

Baselines We compared our method against three baselines: `Baseline`, `FENICE-0.0`, and `FIZZ-0.0`. `Baseline` employs the decoding method described in the Summary Generation section. The latter two only utilize the FENICE and FIZZ metrics, respectively, for reranking.

Table 2 Results for each metric on the CNN/DM dataset. Underline indicates the highest scores for each metric and **Bold** indicates better compared to all baselines. *, †, ‡represent the significance using a bootstrap test against Baseline, FENICE-0.0, and FIZZ-0.0, respectively. "—" indicates removed scores because the reranker maximizes that specific metric, hence they are certainly the best score if included. Each abbreviation represents the following metric, **R1** - ROUGE-1, **R2** - ROUGE-2, **RL** - ROUGE-L, **BS** - BERTScore, **EM** - MENLI-Entailment, **CM** - MENLI-Contradiction, **SM** - MENLI-Summarization, **Fe** - FENICE, and **Fi** - FIZZ.

	Hypotheses as Pseudo-references									Epsilon-Generated Pseudo-references								
	Quality				Factuality					Quality				Factuality				
Method	**R1**	**R2**	**RL**	**BS**	**EM**	**CM**	**SM**	**Fe**	**Fi**	**R1**	**R2**	**RL**	**BS**	**EM**	**CM**	**SM**	**Fe**	**Fi**
Model: BART																		
Baseline	42.28	20.13	35.95	66.53	3.45	-4.10	9.80	98.95	66.36	42.28	20.13	35.95	66.53	3.45	-4.10	9.80	98.95	66.36
FENICE-0.0	42.29	19.41	35.81	66.75	4.99	-4.29	4.61	—	64.17	42.29	19.41	35.81	66.75	4.99	-4.29	4.61	—	64.17
FIZZ-0.0	42.03	19.54	35.72	66.49	4.24	-4.40	5.51	99.19	—	42.03	19.54	35.72	66.49	4.24	-4.40	5.51	99.19	—
FENICE-0.75	**43.27***†‡	**20.44***†‡	**36.52***†‡	**67.12***†‡	3.54	**-3.77***†‡	**13.38***†‡	—	61.38	**43.09***†‡	**20.19**†‡	**36.33***†‡	**67.06***†‡	4.01*	**-3.43***†‡	**13.64***†‡	—	61.91
FIZZ-0.75	**42.74***†‡	**20.18**†‡	**36.24***†‡	**66.83***‡	3.53	**-3.92**†‡	**11.47***†‡	99.15*	—	**42.62***†‡	20.05†‡	**36.10**†‡	**66.79***‡	3.95*	**-3.76***†‡	**11.17***†‡	99.18*	—
MBR-1.0	**43.34***†‡	**20.54***†‡	**36.51***†‡	**67.11***†‡	2.97	**-3.89**†‡	—	98.50	56.47	**43.36***†‡	**20.44***†‡	**36.44***†‡	**67.15***†‡	3.38	**-3.41***†‡	—	98.64	56.62
Oracle	*52.91*	*30.04*	*46.50*	*71.94*	*18.40*	*-0.42*	*37.09*	*99.85*	*89.75*	*52.91*	*30.04*	*46.50*	*71.94*	*18.40*	*-0.42*	*37.09*	*99.85*	*89.75*
Model: Llama-3																		
Baseline	34.83	13.51	28.34	64.04	4.80	-2.12	21.70	98.43	24.98	34.83	13.51	28.34	64.04	4.80	-2.12	21.70	98.43	24.98
FENICE-0.0	35.15	13.77	28.66	64.20	5.53	-2.05	21.19	—	31.12	35.15	13.77	28.66	64.20	5.53	-2.05	21.19	—	31.12
FIZZ-0.0	35.26	13.80	28.74	64.23	5.13	-2.27	20.92	98.89	—	35.26	13.80	28.74	64.23	5.13	-2.27	20.92	98.89	—
FENICE-0.75	34.62	13.68*	28.11	64.16*	4.48	**-1.93***‡	**29.17***†‡	—	26.42*	**35.31***†	**14.11***†‡	28.71*	**64.38***†‡	**5.95***†‡	**-1.60***†‡	**31.14***†‡	—	28.50*
FIZZ-0.75	35.06*	**13.83***	28.52*	**64.24***	4.69	-2.09‡	**24.88***†‡	98.85*	—	**35.36***†‡	**13.98***†‡	**28.80***†‡	**64.34***†‡	5.34*‡	**-1.95***‡	**25.08***†‡	**98.90***	—
MBR-1.0	34.44	13.60*	27.93	64.13*	4.16	**-1.95***‡	—	98.41	23.33	35.15*	**14.07***†‡	28.52*	**64.36***†‡	**6.02***†‡	**-1.51***†‡	—	98.50	24.73
Oracle	*40.88*	*18.83*	*34.05*	*67.03*	*14.81*	*-0.45*	*49.41*	*99.81*	*55.71*	*40.88*	*18.83*	*34.05*	*67.03*	*14.81*	*-0.45*	*49.41*	*99.81*	*55.71*

4 Results

The main results are shown in Table 2. For our proposed method, we tried FENICE and FIZZ with the best weight and only using MBR as the setting. The results span for two conditions, the hypotheses are used as pseudo-references and the pseudo-references are separately generated with increased size for each source document.

Our method **significantly dominate most of the factuality metrics when compared with the baselines in both conditions**. Most MENLI metrics improved significantly, with a slight decrease in the reference-free evaluation scores. Surprisingly, our method is not optimized for Quality metrics, yet most results are significantly better compared to the baselines. These results show that our method works in improving factuality of a summary without reducing the quality of the chosen summary with respect to the "gold" references.

In terms of model type, our method excels in improving both BART and Llama-3's performance. Specifically for BART, our method excels in all of the Quality Metrics and MENLI's score for Factuality metric. Llama-3 has improvements in both Quality and Factuality metrics, specifically BERTScore and MENLI. These results show that our method improves the factuality regardless of the model type.

Comparing the pseudo-references, there is a trade-off in the reference-based and reference-free metrics. All of the reference-based scores slightly decrease, while the reference-free scores increase. This might be due to the number of pseudo-references increases when using Epsilon-Generated condition, reducing the metric bias of MENLI when using MBR. In addition, using Epsilon-Generated condition improves the scores for Llama-3, showing that quality of pseudo-references is important to achieve better consensus.

We calculate Oracle scores for each metric to establish a theoretical upper bound. The differences between the highest scores to the Oracle scores show that there are a big gap of improvement that can be done. Our method is capable of improving the metric by choosing a better summary, but it is not the best summary yet.

5 Conclusion

In this study, we introduced **ConSUM**, a novel reranking method to address the reliance on only the source document or the reference summary as guidance in reranking. Our idea is based on the hypothesis that the best summary is a summary that is not only faithful to the source but also in consensus with the model's distribution. We experimented with two models under two conditions on the CNN/DM dataset. Our findings showed that ConSUM improves summary factuality. Notably, it achieved this without the common trade-off of sacrificing summary quality.

References

[1] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating Summarization Evaluation. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 391–409, 2021.

[2] Jeewoo Sul and Yong Suk Choi. Balancing Lexical and Semantic Quality in Abstractive Summarization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 637–647, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Mathieu Ravaut, Shafiq Joty, and Nancy Chen. SummaReranker: A Multi-Task Mixture-of-Experts Re-ranking Framework for Abstractive Summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4504–4524, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[4] Tanay Dixit, Fei Wang, and Muhao Chen. Improving Factuality of Abstractive Summarization without Sacrificing Summary Quality. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 902–913, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Mathias Müller and Rico Sennrich. Understanding the Properties of Minimum Bayes Risk Decoding in Neural Machine Translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **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)**, pp. 259–272, Online, August 2021. Association for Computational Linguistics.

[6] Bryan Eikema and Wilker Aziz. Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[7] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gu\ensuremath\dotlçehre, and Bing Xiang. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In Stefan Riezler and Yoav Goldberg, editors, **Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning**, pp. 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics.

[8] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024.

[10] Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 4265–4293, Toronto, Canada, July 2023. Association for Computational Linguistics.

[11] Hidetaka Kamigaito, Hiroyuki Deguchi, Yusuke Sakai, Katsuhiko Hayashi, and Taro Watanabe. Diversity explains inference scaling laws: Through a case study of minimum Bayes risk decoding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 29060–29094, Vienna, Austria, July 2025. Association for Computational Linguistics.

[12] Alessandro Scirè, Karim Ghonim, and Roberto Navigli. FENICE: Factuality Evaluation of summarization based on Natural language Inference and Claim Extraction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 14148–14161, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[13] Joonho Yang, Seunghyun Yoon, ByeongJeong Kim, and Hwanhee Lee. FIZZ: Factual Inconsistency Detection by Zoom-in Summary and Zoom-out Document. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 30–45, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

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

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

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

[17] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In **International Conference on Learning Representations**, September 2019.

[18] John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[19] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[20] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In **Eighth International Conference on Learning Representations**, April 2020.

[21] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.

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

A Summarization Models Details

BART BART stands for Bidirectional and Auto-Regressive Transformer [8]. It is a denoising autoencoder that is trained by corrupting text with an arbitrary noising function and learning to reconstruct the original text. Our study used the publicly available `BART-Large` model, fine-tuned to the respective dataset, CNN/DM[1)]

Llama-3 Llama-3 [9] is a family of LLMs developed by Meta. It is a decoder-only transformer that is pre-trained on a massive and diverse dataset. We used the publicly available `Llama-3-8B-Instruct` model[2)]. The prompt used is in Appendix D

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 [19] is an n-gram based metric that is widely used in summarization evaluation. We implement the ROUGE using HuggingFace `evaluate` library.

BERTScore BERTScore [20] is a semantic-based similarity metric. We used the `DeBERTa-XLarge-MNLI` model [22], as it is the top-performing model at the time of this research.[3)]

MENLI MENLI [14] 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.We utilize the three scores provided by MENLI: entailment, contradiction, and summarization. All associated parameters are adopted from the original work.

FENICE FENICE [12] is a two-step factuality metric comprises claim extraction and NLI alignment. We diverge from the original paper's use of ChatGPT in the claim extraction step and instead use the publicly available T5 distillation model provided by the authors[4)].

FIZZ FIZZ [13] is similar to FENICE. However, it provides more interpretability through comparison at atomic fact level. For our FIZZ [13] setup, we use the Orca-2 model[5)] as the decomposer and set the granularity level to 3G.

1) https://huggingface.co/facebook/bart-large-cnn
2) https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
3) Based on the GitHub https://github.com/Tiiiger/bert_score, accessed on July 19, 2025.
4) https://huggingface.co/Babelscape/t5-base-summarization-claim-extractor

Table 3 Average normalized of all metrics for CNN/DM datasets. The 'no mbr' and 'mbr only' settings are denoted by 0 and 1, respectively. Best scores for each reranker are in bold.

w	**0**	**0.25**	**0.50**	**0.75**	**1**
FENICE	27.27	56.60	68.84	**77.71**	63.43
FIZZ	27.27	43.97	58.36	**71.15**	64.68

C Optimal Weight Combination (w)

Experiment Settings This preliminary study aims to find the optimal combination weight (w_{mbr}) between the reference-free scores and MBR scores. We explore 2 different reference-free metrics: FENICE [12] and FIZZ [13]. We generate summaries using validation subset using settings and metrics from §3. We test a range of weights for $w_{mbr} \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, where a weight of 0.0 corresponds to using only the reference-free score and 1.0 corresponds to using only the MBR score.

Results The results of our weight-tuning experiment are presented in Table 3. The table show. To maintain a single, consistent setting, we selected the configuration that performed best across both datasets. We proceed with $w = 0.75$, as it is optimal for CNN/DM and the second-best for XSum. However, due to its negative contribution, we exclude SimCLS from our final system comparison.

D 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}"}
```

5) https://huggingface.co/microsoft/Orca-2-7b