Direct-Scoring NLG Evaluators Can Use Pairwise Comparisons Too

Anonymous EMNLP submission

Abstract

As large-language models have been increasingly used as automatic raters for evaluating free-form content, including document summarization, dialog, and story generation, work has been dedicated to evaluating such models by measuring their correlations with human judgment. For *sample-level* performance, methods which operate by using pairwise comparisons between machine-generated text perform well but often lack the ability to assign absolute scores to individual summaries, an ability crucial for use cases that require thresholding. In this work, we propose a direct-scoring method which uses synthetic summaries to act as pairwise machine rankings at test time. We show that our method performs comparably to state-of-the-art pairwise evaluators in terms of axis-averaged sample-level correlations on the SummEval (**+0.03**), TopicalChat (**-0.03**), and HANNA (**+0.05**) meta-evaluation benchmarks, and release the synthetic in-context summaries as data to facilitate future work.

1 Introduction

As large-language models (LLMs) continue to push the state-of-the-art in natural-language generation (NLG), the task of evaluating the quality of their outputs has become an increasingly complex challenge. Traditional approaches to natural language evaluation that compare n-gram overlap between source and reference text samples (Papineni et al., 2002; Lin, 2004; Denkowski and Lavie, 2014), despite their broad popularity, largely fail to capture semantic information. Model-based evaluation metrics built on smaller pretrained language models (Zhang et al., 2020; Yuan et al., 2021; Zhao et al., 2019; Durmus et al., 2020; Wang et al., 2020; Bhat et al., 2023; Fabbri et al., 2021b,b; Eyal et al., 2019; Chen et al., 2018; Scialom et al., 2021) have offered a more robust solution to NLG evaluation, but often struggle in challenging evaluation scenarios (He et al., 2023; Hanna and Bojar, 2021) and have failed to achieve strong alignment with human judgment on modern meta-evaluation benchmarks (Fabbri et al., 2021a).

To overcome these limitations, recent work has explored the viability of using LLMs for NLG evaluation in a prompting-based scenario, enabling their use as both zero-shot and reference-free evaluators (Liu et al., 2023; Zhou et al., 2024; Liu et al., 2024a). Of these, comparison-based approaches (Gao et al., 2025) have seen considerable attention in recent work, owing largely to their demonstrated superior alignment with human judgment (Liusie et al., 2024).

However, the relative judgment of comparison-based approaches limits their applicability to a number of common use cases, specifically threshold-based scenarios that require an absolute score for filtering and sorting. As an alternative, we propose a direct-scoring evaluator which *uses pairwise comparisons with synthetic samples* and show that it performs comparably to comparison-based approaches over the SummEval (Fabbri et al., 2021a), TopicalChat (Mehri and Eskenazi, 2020), and HANNA (Chhun et al., 2024) meta-evaluation datasets. Finally, we publish all work: the synthetic summaries for each dataset over a variety of LLMs, code and prompts needed to generate the summaries, as well as evaluation utilities.[1]

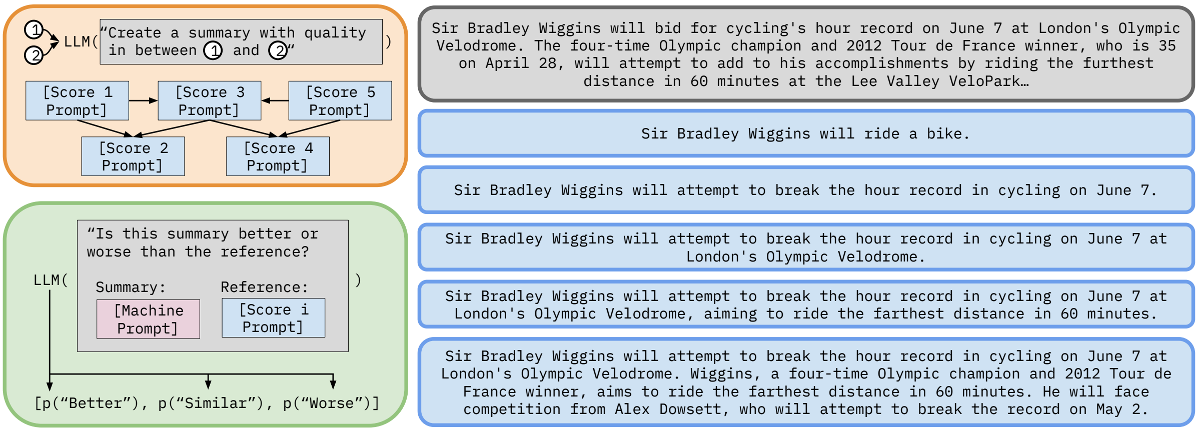
2 Methodology

Next, we describe the proposed method (depicted in Fig. 1) which consists of two pieces: (1) the creation of synthetic in-context examples of various qualities, (2) the inference setup given these generated examples.

2.1 Creating Synthetic In-Context Examples

Given an LLM, task, context, and quality dimension, we wish to generate N responses of varying

[1] https://anonymous.4open.science/r/direct_scoring_synthetic_pairwise-372E/


1
2
LLM("Create a summary with quality in between 1 and 2")
[Score 1 Prompt]
[Score 3 Prompt]
[Score 5 Prompt]
[Score 2 Prompt]
[Score 4 Prompt]
LLM("Is this summary better or worse than the reference?
Summary:
[Machine Prompt]
Reference:
[Score i Prompt])
[p("Better"), p("Similar"), p("Worse")]

Sir Bradley Wiggins will bid for cycling's hour record on June 7 at London's Olympic Velodrome. The four-time Olympic champion and 2012 Tour de France winner, who is 35 on April 28, will attempt to add to his accomplishments by riding the furthest distance in 60 minutes at the Lee Valley VeloPark…

Sir Bradley Wiggins will ride a bike.

Sir Bradley Wiggins will attempt to break the hour record in cycling on June 7.

Sir Bradley Wiggins will attempt to break the hour record in cycling on June 7 at London's Olympic Velodrome.

Sir Bradley Wiggins will attempt to break the hour record in cycling on June 7 at London's Olympic Velodrome, aiming to ride the farthest distance in 60 minutes.

Sir Bradley Wiggins will attempt to break the hour record in cycling on June 7 at London's Olympic Velodrome. Wiggins, a four-time Olympic champion and 2012 Tour de France winner, aims to ride the farthest distance in 60 minutes. He will face competition from Alex Dowsett, who will attempt to break the record on May 2.

Figure 1: **Overview of Method. (Left)** First an LLM is prompted to generate summaries reflecting various levels of quality using a contrastive scheme (orange), then compared to the machine summary to generate probabilities over comparative language (green), eg. "Better", "Similar", and "Worse". **(Right)** We show an example of summaries (blue) of increasing quality (scores 1 through 5) for a SummEval article (grey) on the "consistency" column generated over the course of our method.

quality-levels for the context. For example, we can take the "task" to be summarization, "context" to be news articles, and "responses" to be summaries. Specifically, we want to generate summaries with *inter-rating* consistency: making sure that examples of different scores actually improve with respect to the dimension as rating increases.

We propose to generate examples of monotonically-increasing quality by first starting with the *extremes* of the ratings, then prompting the LLM to generate examples of *intermediate quality* between those previously existing. Specifically, letting $N = 5$ (which matches the rating schemes of SummEval and NewsRoom (Fabbri et al., 2021a; Grusky et al., 2018)) we start by prompting the LLM for the worst and best possible summaries, i.e. scores 1 and 5:

> **(Score 1 - Lowest)** What is the *worst* possible summary of the following article with respect to [quality], [description]?
>
> Article: [article]

> **(Score 5 - Highest)** What is the *best* possible summary of the following article with respect to [quality], [description]?
>
> Article: [article]

where [quality] is the plain-text name of a quality dimension ("Consistency", "Coherence", "Relevance", or "Fluency" in the case of SummEval (Fabbri et al., 2021a)) and [description] is an explanation of [quality]. Let summary$_1$ and summary$_5$ refer to the results of prompting the LLM as above. For the intermediate summaries summary$_{2,3,4}$, we propose to generate them recursively:

> **(Score $i \in 2, 3, 4$ - Intermediate)** For the two summaries, what is a new summary of *intermediate* [quality], [description]?
>
> Article: [article]
> Worse Summary: [summary$_{i-1}$]
> Better Summary: [summary$_{i+1}$]

We provide the final full prompts used for each setting and score in Appendix C and also note that ablation over the prompt and amount of generated examples N is performed in Section 4. From this point, we refer to the summaries as icl$_{1,2,3,4,5}$.

2.2 Pairwise Probability Calculation on Synthetic Examples

Given the synthetic examples icl$_{1,2,3,4,5}$ for a given article and dimension, as well as a prospective machine-generated summary machine, we calculate the probability of the model responding with "Better", "Worse", "Similar", and take the weighted sum over the synthetic scores. Note that this differs from previous direct scoring approaches like G-Eval (Liu et al., 2023), which take the probabilities over scores themselves; we prompt each score *pairwise* with a generated reference, then take the summation.

> **(Score i Evaluation)** Here's a news article, a reference summary, and prospective summary. How does the prospective summary compare to the reference with respect to [quality], [description]? Respond only with "Worse," "Better," or "Similar."
>
> Article: [article]
> Reference Summary: [icl$_i$]
> Prospective Summary: [machine]

	SummEval					TopicalChat				HANNA			
Method	**COH**	**CON**	**FLU**	**REL**	**AVG**	**NAT**	**ENG**	**OVE**	**AVG**	**COH**	**SUR**	**COM**	**AVG**
Other Metrics													
BERTScore[1] (F1)	0.28	0.19	0.11	**0.31**	0.23	**0.09**	**0.16**	**0.19**	**0.15**	**0.25**	0.22	**0.30**	**0.26**
BERTScore[2] (F1)	**0.38**	**0.35**	**0.40**	0.28	**0.35**	0.04	-0.04	0.07	0.02	**0.25**	0.12	0.17	0.18
GPTScore	0.28	0.31	0.38	0.22	0.30	-	-	-	-	0.22	**0.25**	0.08	0.18
Mistral 7B													
ZEPO	0.29	0.32	0.13	0.30	0.26	0.14	0.25	0.28	0.22	-	-	-	-
PairS-beam	0.28	0.31	0.18	0.24	0.25	**0.41**	**0.41**	0.33	**0.38**	0.29	**0.27**	0.31	0.29
Direct Scoring	0.23	0.37	0.19	0.19	0.25	0.26	0.17	0.32	0.25	0.30	0.26	0.37	0.31
G-Eval	0.25	**0.39**	**0.20**	0.25	0.27	0.26	0.28	**0.35**	0.30	**0.34**	0.25	0.39	0.33
Ours	**0.35**	0.34	0.18	**0.33**	**0.30**	0.31	0.36	0.22	0.30	0.33	0.25	**0.44**	**0.34**
Llama-3 8B													
ZEPO	0.40	0.25	0.30	0.39	0.34	0.16	0.26	0.46	0.29	-	-	-	-
PairS-beam	0.35	**0.42**	**0.32**	0.35	0.36	**0.47**	**0.56**	0.43	**0.49**	0.36	0.22	0.31	0.30
Direct Scoring	0.35	0.32	0.23	**0.46**	0.34	0.33	0.32	0.40	0.35	0.26	0.17	0.32	0.25
G-Eval	0.34	0.29	0.22	0.42	0.32	0.38	0.43	**0.53**	0.45	**0.44**	0.29	0.42	0.38
Ours	**0.44**	0.41	0.26	0.36	**0.37**	0.43	0.50	0.43	0.46	0.42	**0.37**	**0.49**	**0.43**

Table 1: **Main Results.** Sample-level Spearman correlations of Llama-3 8B over the various axes of SummEval (Fabbri et al., 2021a), TopicalChat (Mehri and Eskenazi, 2020), and HANNA (Chhun et al., 2024) for comparison-based evaluators and direct-scoring evaluators. "Average" refers to the mean of the columns for each dataset. Best performance is **bolded** and second-best is underlined. BERTScore[1] refers to using `bert-base-uncased` whereas [2] refers to using `roberta-large`.

Given the above prompt for score i, we calculate and sum the log probabilities of the LLM responding with `"Worse"`, `"Better"`, or `"Similar"`. Next, we calculate the softmax over these summed log probabilities, which we refer to as $p(\texttt{"Worse"}|i)$, $p(\texttt{"Better"}|i)$, and $p(\texttt{"Similar"}|i)$. Using these probabilities, we construct final prediction by taking weighted average of scores:

$$s(\cdot) = \sum_{i \in 1,\dots,5} [i, -i, 0] * \begin{bmatrix} p(\texttt{"Better"}|i) \\ p(\texttt{"Worse"}|i) \\ p(\texttt{"Similar"}|i) \end{bmatrix} \quad 135$$

3 Experiments

Datasets We evaluate our method on three meta-evaluation benchmarks: SummEval (Fabbri et al., 2021a) for summarization, TopicalChat (Mehri and Eskenazi, 2020) for dialog, and HANNA (Chhun et al., 2024) for story generation. Contrary to prior works, we do not report NewsRoom (Grusky et al., 2018) in the main results due to a surprising finding that BERTScore (Zhang et al., 2020) with `roberta-large` attains high performance (see Appendix A).

Metric Following previous works (Zhou et al., 2024; Zhong et al., 2022; Fu et al., 2023), we evaluate the efficacy of our method using *sample-level* correlations with machine generated summaries. Namely, for a given correlation metric (e.g. Spearmans ρ) and axis, a group of abstractive summarization machines of size M, and a number of full-texts T, we calculate the following metric:

$$F^{sample} = \frac{1}{n} \sum_{i=1}^{T} \rho(\begin{bmatrix} pred_{i,1}, human_{i,1} \\ \dots \\ pred_{i,M}, human_{i,M} \end{bmatrix}) \quad 155$$

where $pred_{i,j}$ is the prediction for the j^{th} machine's response on the i^{th} document and $human_{i,j}$ is the ground truth label for that machine response. We provide a brief discussion of the aggregation metrics, namely the originally proposed *system-level* (Fabbri et al., 2021a) and *summary-level* metrics in Appendix B.

Baselines We provide direct-scoring baselines G-Eval (Liu et al., 2023), a direct-scoring baseline ("Scoring") which uses the same prompt as G-Eval, as well as classical metrics such as BERTScore (Zhang et al., 2020) and GPTScore (Fu et al., 2023). For pairwise baselines, we provide ZEPO (Zhou et al., 2024) and PairS-beam (Liu et al., 2024a). For all sampling-based generation, we conduct all experiments with the same hyperparameters.

4 Results

Main Results We show the sample-level performance of our method in Table 1. In terms of average sample-level correlation across the axes of each dataset, our method performs the best on SummEval (+0.03/+0.01 for Mistral 7B (Jiang et al., 2023) / Llama-3 8B (Grattafiori et al., 2024) versus next highest performer), second best on TopicalChat (**-0.08/-0.03**), and best on HANNA (**+0.01/+0.05**). However, we do not find that our method is always better over axes, often being the first or runner-up.

Performance of Different Prediction Methods In the first section of Table 2, we ablate on the prediction method used to generate final scores for a given text. "Sample" refers to sampling n responses using constrained decoding and "$p(\texttt{"Yes"}), p(\texttt{"No"})$" refers to using $\texttt{"Yes"}$ and $\texttt{"No"}$ with a comparison-based prompt (see Appendix C). Firstly, we find that using a comparative-based prompt in our setup increases performance over the proposed setting. Next, we find that increasing n results in monotonically increasing performance, with the maximum achieved at $n = 1000$ with **36.44** sample-level correlation. We also note that this is still below the performance of our method (**37.43** - $p(\texttt{"Better"}), p(\texttt{"Worse"}), \ldots$), notably a gap of **0.99**. This indicates that probability generation is essential to our method, which we fix for the second section of the table.

When varying the amount of examples to compare to at prediction time (second section of Table 2), we see that increasing the amount of examples, even past $N = 5$, increases performance with the maximum achieved by $N = 9$ (**37.80** vs. **37.43**). However, we still keep $N = 5$ for simplicity and its alignment with popular meta-evaluation datasets.

Method	Avg. SummEval
Sampling	
Sample, $n = 1$	9.38 ± 0.67
Sample, $n = 3$	14.68 ± 1.51
Sample, $n = 10$	21.24 ± 2.26
Sample, $n = 100$	30.91 ± 1.32
Sample, $n = 1000$	36.44 ± 0.42
$p(\texttt{"Better"}), p(\texttt{"Worse"}), \ldots$	37.43
$p(\texttt{"Yes"}), p(\texttt{"No"}), \ldots$	**38.60**
# of Examples	
$N = 2$	34.83
$N = 3$	36.58
$N = 5$	37.43
$N = 9$	**37.80**

Table 2: **Ablation on Prediction Method Used.** "Sampling" refers to having the LLM generate a pairwise judgment via constrained decoding, whereas N refers to the amount of in-context examples used to compare to the target summary. "±" refers to the standard deviation over 5 trials.

Method	MMLU	Direct	Ours
Prediction LLM			
Mistral-v0.1-7B	60.1	0.20	0.30
OLMo-2-7B	61.3	0.05	0.32
Llama-3-8B	66.6	**0.33**	0.37
Llama-3.1-8B	**71.3**	0.32	**0.45**
Examples LLM			
Mistral-v0.1-7B	60.1	0.20	0.22
OLMo-2-7B	61.3	0.05	**0.20**
Llama-3-8B	66.6	**0.33**	0.37
Llama-3.1-8B	**71.3**	0.32	**0.42**

Table 3: **Ablation on LLMs Used for Example Generation and Prediction.** Methods are sorted by ascending MMLU (5-shot) (Hendrycks et al., 2020). All models are their "Instruct" variant. Prompts used for each result are detailed in Appendix C.

Performance Across Architectures In the first section of Table 3, we show the performance of our method when using different LLMs for prediction after generating synthetic examples. We find that the performance of our method can be further boosted by using more powerful backbones, as in the vanilla Direct Scoring setting, models with greater all-around performance (as measured by MMLU ↑) do not correspond to better-performing evaluators, but under our method this discrepancy is fixed.

Similarly, when using different LLMs for synthetic example generation (second section of Table 3), we find performance increases with more powerful models, albeit less consistently. Most notably, OLMo-2-7B (OLMo et al., 2024) underperforms relative to its MMLU score, producing less useful summaries than Mistral-v0.1-7B.

5 Conclusion

In this work, we proposed an LLM-based direct-scoring evaluation framework which uses synthetic in-context examples from LLMs to assign absolute scores to machine-generated summaries. We found that the method produces comparable average sample-level correlations to comparison-based approaches. We ablated on the method to find that probability generation is essential to performance, increasing the granularity of examples moderately boosts performance, and that LLMs with more instruction-following ability are higher performing with our method. This method addresses the need for a direct scoring metric with performance comparable to that of state-of-the-art comparison-based approaches, allowing for use cases involving thresholding. We also publicly release the synthetic summaries for further work.

6 Limitations

Choice of LLMs We ablate on on a selection of small-sized (between 7 and 8 billion parameters) LLMs to compare to prior work and to adhere to our computational budget. Behavior for this subset of models may not extrapolate to larger LLMs. We leave investigations into the scalability of our approach with regard to model size to future work.

Computational Cost Our proposed approach requires N synthetic reference examples to be generated per source article and task (quality dimension), plus an additional N inputs to the LLM to compare the input text to each of the reference examples. This could become computationally prohibitive for scenarios in which there is no repetition in source articles, especially for larger values of N, resulting in both reduced inference speed and higher financial cost. Given that our ablation studies (see Table 2) show reduced performance at lower values of N, future work into both increasing synthetic summary generation efficiency and improving performance at lower values of N may be useful.

References

Meghana Moorthy Bhat, Rui Meng, Ye Liu, Yingbo Zhou, and Semih Yavuz. 2023. Investigating answerability of llms for long-form question answering. *arXiv preprint arXiv:2309.08210*.

Ping Chen, Fei Wu, Tong Wang, and Wei Ding. 2018. A semantic qa-based approach for text summarization evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Cyril Chhun, Fabian M Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. *arXiv preprint arXiv:1906.00318*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021b. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–28.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024a. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulic, Anna Korhonen, and Nigel Collier. 2024b. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

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

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

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 multidimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Han Zhou, Xingchen Wan, Yinhong Liu, Nigel Collier, Ivan Vulić, and Anna Korhonen. 2024. Fairer preferences elicit improved human-aligned large language model judgments. *arXiv preprint arXiv:2406.11370*.

A BERTScore and NewsRoom

Method	**COH**	**REL**	**INF**	**FLU**	**AVG**
Other Metrics					
BERTScore[1] (F1)	0.15	0.16	0.13	0.17	0.15
BERTScore[2] (F1)	**0.65**	**0.62**	**0.69**	**0.58**	**0.63**
GPTScore	0.31	0.35	0.26	0.31	0.31
Mistral 7B					
ZEPO	0.47	0.38	0.44	0.48	0.44
PairS-beam	**0.55**	**0.53**	**0.48**	**0.48**	**0.51**
Direct Scoring	0.32	0.39	0.20	0.26	0.29
G-Eval	0.36	0.36	0.24	0.39	0.34
Llama-3 8B					
ZEPO	0.57	0.54	0.55	0.56	0.56
PairS-beam	**0.66**	**0.66**	**0.73**	**0.62**	**0.67**
Direct Scoring	0.42	0.41	0.30	0.29	0.36
G-Eval	0.38	0.34	0.26	0.26	0.31
GPT-4-Turbo					
ZEPO	-	-	-	-	-
PairS-beam	**0.64**	**0.61**	**0.67**	**0.60**	**0.63**
Direct Scoring	0.55	0.54	0.57	0.60	0.57
G-Eval	0.58	0.55	0.57	0.58	0.57

Table 4: **NewsRoom Performance.** BERTScore[1] refers to using `bert-base-uncased` whereas [2] refers to using `roberta-large`. The performance of `roberta-large` is highlighted in red. Other numbers are reported from PairS (Liu et al., 2024a) and ZEPO (Zhou et al., 2024)

The NewsRoom dataset is often cited in works involving summarization metrics. However, after computing the BERTScore for the subset of NewsRoom's test data that includes human evaluations, we found that BERTScore with `roberta-large` attained results comparable to those of a direct scoring approach with GPT-4-turbo (Grusky et al., 2018; Liu et al., 2024b). Conversely, BERTScore with roberta-large performed markedly worse than GPT-4-turbo on the SummEval annotated dataset, presenting an inconsistency in the expected discrepancies between the results of the methods tested on NewsRoom (Liu et al., 2024b).

This is supported by NewsRoom's overall higher correlation between axes, displayed in Table 6, which could result in unusually high scores from BERTScore due to the lack of discernment between the axes in the provided ground-truth human evaluation scores. For comparison, we also include SummEval axis correlations in Table 5, whose correlations with the exception of the coherence/relevance pair are lower.

	COH	FLU	CON	REL
COH	1.000	0.197	0.400	0.787
FLU	0.197	1.000	0.317	0.254
CON	0.400	0.317	1.000	0.458
REL	0.787	0.254	0.458	1.000

Table 5: **SummEval Spearman Axis Sample-Level Correlations**. COH, FLU, CON, and REL are the abbreviations of "coherence," "fluency," "consistency," and "relevance," respectively.

	COH	FLU	INF	REL
COH	1.000	0.788	0.753	0.559
FLU	0.788	1.000	0.674	0.547
INF	0.753	0.674	1.000	0.624
REL	0.559	0.547	0.624	1.000

Table 6: **NewsRoom Spearman Axis Sample-Level Correlations**. COH, FLU, INF, and REL are the abbreviations of "coherence," "fluency," "informativeness," and "relevance," respectively.

B System, Sample, and Summary-Level Performance Definitions

There are three main correlation metrics found within NLG meta-evaluation works: *system-level*, *sample-level*, and *summary-level*. In terms of the SummEval (Fabbri et al., 2021a) dataset, these correspond to (1) the correlation of the average machine score with the automatic rater versus the human score, (2) the average machine correlation within each document, and (3) the average correlation over all documents after throwing away the machine ID. Formally, for a given correlation metric (e.g. Spearman's ρ), a group of abstractive summarization machines of size M, and a number of full-texts T, we can choose the following metrics:

System-Level:

$$F^{system} = \rho(\frac{1}{n}\sum_{i=1}^{T}\begin{bmatrix} pred_{i,1}, & human_{i,1} \\ & ... \\ pred_{i,M}, & human_{i,M} \end{bmatrix})$$

Sample-Level:

$$F^{sample} = \frac{1}{n}\sum_{i=1}^{T}\rho(\begin{bmatrix} pred_{i,1}, human_{i,1} \\ ... \\ pred_{i,M}, human_{i,M} \end{bmatrix})$$

Summary-Level:

$$F^{summary} = \rho(\begin{bmatrix} pred_{1,1}, human_{1,1} \\ ... \\ pred_{1,M}, human_{1,M} \\ pred_{2,1}, human_{2,1} \\ ... \\ pred_{2,M}, human_{2,M} \\ ... \\ pred_{T,1}, human_{T,1} \\ ... \\ pred_{T,M}, human_{T,M} \end{bmatrix})$$

where $pred_{i,j}$ is the prediction for the j^{th} machine's response on the i^{th} document and $human_{i,j}$ is the ground truth label for that machine response.

C Prompts Used for Summary Generation and Prediction

Next, we provide the templates for generating synthetic summaries and prediction for different schemes. We organize the section in the following way:

1. SummEval
   (a) Best/Worst (Scores 1,5): Table 7
   (b) Recursive (Scores 2,3,4): Table 10
   (c) $p($"Better"|i$)$, ...: Table 13
   (d) $p($"Yes"|i$), p($"No"|i$)$: Table 16
2. TopicalChat
   (a) Best/Worst (Scores 1,5): Table 8
   (b) Recursive (Scores 2,3,4): Table 11
   (c) $p($"Better"|i$)$, ...: Table 14

(d) $p($"Yes"$|$i$), p($"No"$|$i$)$: Table 17

3. HANNA
   (a) Best/Worst (Scores 1,5): Table 9
   (b) Recursive (Scores 2,3,4): Table 12
   (c) $p($"Better"$|$i$), ...$: Table 15
   (d) $p($"Yes"$|$i$), p($"No"$|$i$)$: Table 18

```
You will be given a source document and an evaluation dimension for a summary. Your
task is to write the {{ worst_best : str } possible summary you can think of with
regards to this dimension.

Your response should only include the {{ worst_best : str }} possible summary you
can create without any additional text. The summary must be non-empty and directly
summarize the article without using phrases like ”This article is about.”

Evaluation Criteria:

{{ col_title : str }} - {{ col_description : str }}

Document:

{{ article : str }}
```

Table 7: **SummEval Best/Worst Synthetic Example Prompt Template.**

```
You will be given a conversation between two people and an evaluation dimension for
a response to the most recent message. Your task is to write the {{ worst_best : str
}} possible response you can think of with regards to this dimension.

Your response should exactly be the {{ worst_best : str }} possible response without
any additional text. The response must be non-empty. Try to make the response less
than a few sentences and keep the tone informal.

Evaluation Criteria:
{{ col_title : str }} - {{ col_description : str }}

Conversation:
{{ context : str }}
```

Table 8: **TopicalChat Best/Worst Synthetic Example Prompt Template.**

```
You will be given an idea for a story and an evaluation dimension for that story.
Your task is to write the {{ worst_best : str }} possible story you can think of
with regards to this dimension.

Your response should exactly be the {{ worst_best : str }} possible story without any
additional text. The summary must be non-empty and directly summarize the article
without using phrases like ”This story is about.” Try to keep the story less than
150 words and end on a full sentence. Don’t use paragraphs.

Evaluation Criteria:
{{ col_title : str }} - {{ col_description : str }}

Story Idea:
{{ story_prompt : str }}
```

Table 9: **HANNA Best/Worst Synthetic Example Prompt Template.**

```
You will be given a source document and an evaluation dimension for a summary. Your task is to write a
summary which is higher quality than one summary (Bad Summary) but lower quality than another (Good
Summary).

Your response should only include the in-between summary without any additional text. The summary
must be non-empty and directly summarize the article without using phrases like "This article is about."

Evaluation Criteria:
{{ col_title : str }} - {{ col_description : str }}

Bad Summary:
{{ worse_summary : str }}

Good Summary:
{{ better_summary : str }}

Document:
{{ article : str }}
```

Table 10: **SummEval Recursive Synthetic Example Prompt Template.**

```
You will be given a conversation between two people and an evaluation dimension for a response to the
most recent message. Your task is to write a response which is higher quality than one response (Bad
Response) but lower quality than another (Good Response).

Your message should exactly be the in-between response without any additional text. The response must
be non-empty. Try to make the response less than a few sentences and keep the tone informal.

Evaluation Criteria:
{{ col_title : str }} - {{ col_description : str }}

Conversation:
{{ context : str }}

Bad Response:
{{ worse_summary : str }}

Good Response:
{{ better_summary : str }}
```

Table 11: **TopicalChat Recursive Synthetic Example Prompt Template.**

```
You will be given an idea for a story and an evaluation dimension for a story. Your task is to write a
story which is higher quality than one story (Bad Story) but lower quality than another (Good Story).

Your response should exactly be the in-between story without any additional text. The story must be
non-empty, be related to the idea, and not use phrases like "This story is about." Try to keep the
story less than 150 words and end on a full sentence. Do not use paragraphs.

Evaluation Criteria:
{{ col_title : str }} - {{ col_description : str }}

Story Idea:
{{ story_prompt : str }}

Bad Story:
{{ worse_summary : str }}

Good Story:
{{ better_summary : str }}
```

Table 12: **HANNA Recursive Synthetic Example Prompt Template.**

```
You will be given a news article, one target summary of that news article, and a reference summary of that article. Your goal is
to say whether the quality of the target summary is better, worse, or similar to the reference summary with respect to {{ col : str }}.

Evaluation Criteria:
{{ col_title }}: {{ col_description }}

Evaluation Steps:
1. Read the news article carefully and identify the main facts and details it presents.
2. Read the target summary and example summary. Compare them to the article.
3. Compare the quality of the target summary to reference summary with respect to {{ col : str }}.
4. Respond with only one of the following: "Better" "Worse" or "Similar" which indicate whether the target summary is better than,
worse than, or similar to the reference summary.

Original Article:
{{ article : str }}

Reference Summary:
{{ icl_summary : str }}

Target Summary:
{{ target_summary : str }}
```

Table 13: **SummEval** $p(\texttt{"Better"})$, ... **Prediction Prompt Template.**

```
You will be given a news article, one target summary of that news article, and a reference summary of that article. Your goal is
to say whether the quality of the target summary is better, worse, or similar to the reference summary with respect to {{ col : str }}.

Evaluation Criteria:
{{ col_title }}: {{ col_description }}

Evaluation Steps:
1. Read the story carefully.
2. Read the reference story and evaluation story. Compare them to the idea.
3. Compare the quality of the evaluation story to the reference story with respect to {{ col : str }}.
4. Respond with only one of the following: "Better" "Worse" or "Similar" which indicate whether the evaluation story is better than,
worse than, or similar to the reference story.

Story Idea:
{{ story_prompt : str }}

Reference Story:
{{ icl_summary : str }}

Evaluation Story:
{{ target_summary : str }}
```

Table 14: **TopicalChat** $p(\texttt{"Better"})$, ... **Prediction Prompt Template.**

```
You will be given a news article, one target summary of that news article, and a reference summary of that article. Your goal is
to say whether the quality of the target summary is better, worse, or similar to the reference summary with respect to {{ col : str }}.

Evaluation Criteria:
{{ col_title }}: {{ col_description }}

Evaluation Steps:
1. Read the story carefully.
2. Read the reference story and evaluation story. Compare them to the idea.
3. Compare the quality of the evaluation story to the reference story with respect to {{ col : str }}.
4. Respond with only one of the following: "Better" "Worse" or "Similar" which indicate whether the evaluation story is better than,
worse than, or similar to the reference story.

Story Idea:
{{ story_prompt : str }}

Reference Story:
{{ icl_summary : str }}

Evaluation Story:
{{ target_summary : str }}
```

Table 15: **HANNA** $p(\texttt{"Better"})$, ... **Prediction Prompt Template.**

```
Here is a news article:
{{ article : str }}

Summary 1:
{{ target_summary : str }}

Summary 2:
{{ icl_summary : str }}

Does Summary 1 {{ prediction : str }} than Summary 2?

Respond with only one of the following: "Yes" or "No".
```

Table 16: **SummEval** $p(\texttt{"Yes"}), p(\texttt{"No"})$ **Prediction Prompt Template.**

```
Here is a conversation:
{{ context : str }}

Response 1:
{{ target_summary : str }}

Response 2:
{{ icl_summary : str }}

Does Response 1 {{ prediction : str }} than Response 2?

Respond with only one of the following: "Yes" or "No".
```

Table 17: **TopicalChat** $p(\texttt{"Yes"}), p(\texttt{"No"})$ **Prediction Prompt Template.**

```
Story 1:
{{ target_summary : str }}

Story 2:
{{ icl_summary : str }}

Does Story 1 {{ prediction : str }} than Story 2?

Respond with only one of the following: "Yes" or "No".
```

Table 18: **HANNA** $p(\texttt{"Yes"}), p(\texttt{"No"})$ **Prediction Prompt Template.**