Multi-Dimensional Evaluation of Text Summarization with In-Context Learning

Sameer Jain[1] **Vaishakh Keshava**[1] **Swarnashree Mysore Sathyendra**[1]
Patrick Fernandes[1,2] **Pengfei Liu**[1] **Graham Neubig**[1] **Chunting Zhou**[3]
[1]Carnegie Mellon University [2]Instituto Superior Técnico [3]Facebook AI Research
{sameerj, vkeshava, smysores}@cs.cmu.edu

Abstract

Evaluation of natural language generation (NLG) is complex and multi-dimensional. Generated text can be evaluated for fluency, coherence, factuality, or any other dimensions of interest. Most frameworks that perform such multi-dimensional evaluation require training on large manually or synthetically generated datasets. In this paper, we study the efficacy of large language models as multi-dimensional evaluators using in-context learning, obviating the need for large training datasets. Our experiments show that in-context learning-based evaluators are competitive with learned evaluation frameworks for the task of text summarization, establishing state-of-the-art on dimensions such as relevance and factual consistency. We then analyze the effects of factors such as the selection and number of in-context examples on performance. Finally, we study the efficacy of in-context learning-based evaluators in evaluating zero-shot summaries written by large language models such as GPT-3. Our code is available at https://github.com/JainSameer06/ICE

1 Introduction

Developing comprehensive evaluation frameworks (Deng et al., 2021; Yuan et al., 2021; Zhong et al., 2022) that can evaluate multiple human-interpretable dimensions, such as factual consistency (Kryscinski et al., 2020; Wang et al., 2020) and coherence (Dziri et al., 2019; Huang et al., 2020), is important for the advancement of Natural Language Generation (NLG). However, similarity-based metrics (Papineni et al., 2002; Lin, 2004; Sellam et al., 2020; Zhao et al., 2019; Zhang et al., 2020) still dominate NLG evaluation in practice. Compared to them, desired multi-dimensional evaluators do not require reference texts for evaluation; and they can easily extend to new explainable evaluation dimensions. Recently, Zhong et al. (2022) developed a unified evaluation framework that can generalize to multiple dimensions and text generation tasks. However, it relies on the construction of synthetic and auxiliary data for the finetuning of a pre-trained language model, requiring in-depth knowledge and significant engineering effort for each dimension. Furthermore, the inclusion of new dimensions requires (continued) training of the model, and might affect the performance on other dimensions in unforeseen ways.

```
Text: Cats and dogs have the advantage...
Summary: roland girous keeps a blood parrot...
Consistency: 1.0

Text: Jordan Henderson has provided Liverpool..
Summary: jordan henderson is set to sign a...
Consistency: 0.67

Text: Arsenal playmaker Mesut Ozil seemed...
Summary: mesut ozil impressed on international...
Consistency: _____
```

Figure 1: Our prompt design to evaluate the consistency of the summary in red, illustrated using two in-context examples (in blue). To evaluate other aspects, we remove the source text or replace it with a reference.

In this work, we propose to use *in-context* learning (Brown et al., 2020) with large language models (LLMs) — a commonly used method to perform many tasks by utilizing only a few input-output examples — to perform multi-dimensional text evaluation in a unified fashion. Compared to pre-trained evaluators that need specialized supervised training for each dimension, our **I**n-**C**ontext learning-based **E**valuator (ICE) framework is:

- Learning-free. It does not require supervised fine-tuning on large annotated (synthetic) training data, requiring only a handful of samples at inference time.
- Extensible. To evaluate new dimensions, it does not rely on large amounts of human judgments or the construction of new synthetic data, using only a natural language prompt consisting of a small number of example pairs to ascertain the properties associated with a given quality aspect.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8487–8495
July 9-14, 2023

In this paper, using text summarization as a test bed, we show that with a simple prompt design, ICE is competitive with state-of-the-art trained evaluators on multi-dimensional evaluation of model-produced summaries, establishing a new state-of-the-art on dimensions such as relevance and factual consistency. To study the robustness of the evaluator to the selection of in-context examples, we analyze the factors that affect the performance of ICE, such as the number of in-context examples and sampling procedures when picking in-context examples from a set of candidates. We find ICE to be robust to the selection of in-context examples and observe a slight improvement in performance as the number of examples is increased. Finally, in light of the recent work (Goyal et al., 2022) that points to the misalignment of existing evaluation metrics with human preference in evaluating zero-shot summaries generated by LLMs such as GPT-3 (Brown et al., 2020), we study the effectiveness of ICE in evaluating zero-shot summaries generated by GPT-3. We find that ICE evaluations agree closely with human judgments on such summaries.

2 Methodology

2.1 Problem Statement

Given a sequence $\mathbf{x}$ that is input to an NLG system and a system-generated output sequence $\mathbf{y}$, an evaluation framework outputs a score s that captures the quality of $\mathbf{y}$, either with or without the help of a human-generated reference output $\mathbf{r}$.[1] In case of multi-dimensional evaluation where we are interested in assessing $\mathbf{y}$ over d quality metrics, we instead get a vector $\mathbf{S} = (s_1, s_2, ..., s_d)$ over diverse dimensions (e.g., coherence, fluency). Depending on the dimension, there is sometimes a need to condition an evaluation on $\mathbf{x}$ (such as to evaluate consistency in summarization). We evaluate our method over four dimensions:

- Consistency: The factual correctness of a summary given the source text.
- Relevance: The property of capturing salient information from the source.
- Fluency: A measure of the quality of the individual sentences in the summary.
- Coherence: A measure of the quality, organization, and structure of sentences in the summary.

[1] Specifically for summarization, most learned frameworks evaluate relevance through reference-based evaluation.

2.2 Prompt Design & Score Extraction

ICE relies on an LLM (we use the `text-davinci-003` model of GPT-3) to make predictions. It takes in a prompt that consists of a small number of in-context examples, each of which consists of generated text and its corresponding quality score as a numeric string. The prompt ends with a test example, for which the model predicts a score (Figure 1).

The input contains the model-generated text (summary), in addition to which it might contain additional information such as the source text or references, depending on the dimension. To evaluate `fluency` and `coherence`, our prompts use in-context examples consisting of generated summaries and corresponding scores. For `consistency` and `relevance`, we use the source text and a reference summary respectively, in addition to the generated summary. We pass this prompt to a GPT-3 model, with sampling temperature set to 0 to elicit deterministic responses. We parse the model response–decoded numeric string–as the dimension score.

2.3 Selection of In-context Examples

By default, we use 4 in-context examples in our prompts, as this is the largest number that fits within the context window of GPT-3. We experiment with two sampling procedures (Appendix B) to obtain 4 examples from a pool of examples:

1. **Uniform Random Sampling**. We randomly select 4 summaries from the pool of examples. This causes the examples to follow the same distribution as the example pool.
2. **Stratified Sampling**. We bucket the range of scores, i.e. $[0, 1]$, into 4 equal partitions and randomly sample one summary from each one. This causes examples to be representative of the range of scores in the example pool.

We avoid using synthetically generated data (Kryscinski et al., 2020; Zhong et al., 2022) since the kind of errors made by generation models is often different from the errors present in the negative examples in these datasets (Goyal and Durrett, 2021). We instead elect to use (a few) human evaluations of model-generated text in order to make the in-context examples as representative of real errors as possible. We do this by splitting the meta-evaluation dataset and using a partition as an in-context example pool, as described in Section 3.1.

Metric	Coherence		Consistency		Fluency		Relevance	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
CTC	-	-	0.425	0.340	-	-	**0.495**	0.364
BARTScore	0.445	0.340	0.380	0.314	0.345	0.283	0.357	0.274
UniEval	**0.591**	**0.424**	0.433	0.348	**0.445**	**0.349**	0.473	0.343
ICE (Uniform Sampling)	0.476	0.388	**0.486**	**0.466**	0.366	0.328	0.467	0.384
ICE (Stratified Sampling)	0.497	0.387	0.298	0.263	0.397	0.348	0.485	**0.396**

Table 1: Summary-level Spearman and Kendall-Tau correlations of different metrics on the SummEval benchmark

3 Experiments

3.1 Datasets & Baselines

We use the SummEval dataset (Fabbri et al., 2020)[2] to meta-evaluate our evaluation framework. SummEval collects human evaluation annotations for 16 summarization systems on 100 articles sampled from the CNN/DailyMail corpus, for a total of 1600 summary-level annotations. Each summary is evaluated on four dimensions described in Section 2.2.

To get a pool of in-context examples, we keep aside a small subset (64 examples) of the SummEval dataset to pick in-context examples from, and use the rest (1536 examples) as the test set for meta-evaluation (evaluating the baselines on this same test set). Further details are in Appendix A.

We compare ICE to the following state-of-the-art multi-dimensional evaluators: (1) **CTC** (Deng et al., 2021) uses information alignment between generated outputs and references or inputs; (2) **BARTScore** (Yuan et al., 2021) uses the conditional probability of a sequence given inputs or references; and (3) **UniEval** (Zhong et al., 2022) uses a question-answering framework (e.g. *"Is this a coherent summary?"*) to calculate metrics.

Following Liu et al. (2021); Zhong et al. (2022), we assess performance by computing summary-level Spearman and Kendall-Tau correlations between predicted scores and human judgements.

3.2 Results

As illustrated in Table 1, ICE is competitive with fine-tuned baselines despite not requiring any finetuning. It achieves state-of-the-art correlation with human judgments for `relevance` and `consistency`. We perform pairwise significance tests and observe that ICE (uniform sampling) does better than UniEval on `consistency` and `relevance` on Kendall's Tau with a significance level of 0.05 (Appendix E). Additionally, the uniform sampling variant of ICE outperforms BARTScore (which also does not require fine-tuning) across dimensions.

Between the two *sampling procedures* for ICE, we observe that stratified sampling works marginally better for all dimensions other than `consistency`. Since summaries in the SummEval dataset have perfect or near-perfect human scores for `consistency` (Figure 2), uniform sampling causes in-context examples to also have near-perfect scores. This appears useful for the model to calibrate its scoring when evaluating `consistency`, leading to better performance. We explore this in greater detail in §4.1. While the same reasoning could hold for `fluency`, we observe both here and in §4.3 that `fluency` scores are quite stable. Given that `fluency` is an easier aspect to evaluate, this stability could be a result of the model possessing a strong notion about `fluency` from pre-training time that is not modified significantly as the distribution of in-context examples changes (Reynolds and McDonell, 2021). Finally, we observe that the performance for `coherence` and `relevance` are similar regardless of the sampling procedure. This is because scores for these aspects are spread out in the dataset, which makes uniform and stratified sampling return similar in-context examples.

4 Analysis

In this section, we analyse the effects of our prompt engineering choices. The comparison between sampling procedures in Section 4.1 is performed on the entire test set but the experiments in Sections 4.2 and 4.3 are performed on a test set sample of size 200 to control costs. The analyses in Sections 4.1 and 4.2 use four in-context examples.

4.1 Analyzing the Sampling Procedures

Figure 2 illustrates that the prediction distributions from uniform and stratified sampling differ the most when the true distribution is skewed, such as for `consistency`. In such a case, stratified sampling selects in-context examples from the entire

[2] https://github.com/Yale-LILY/SummEval

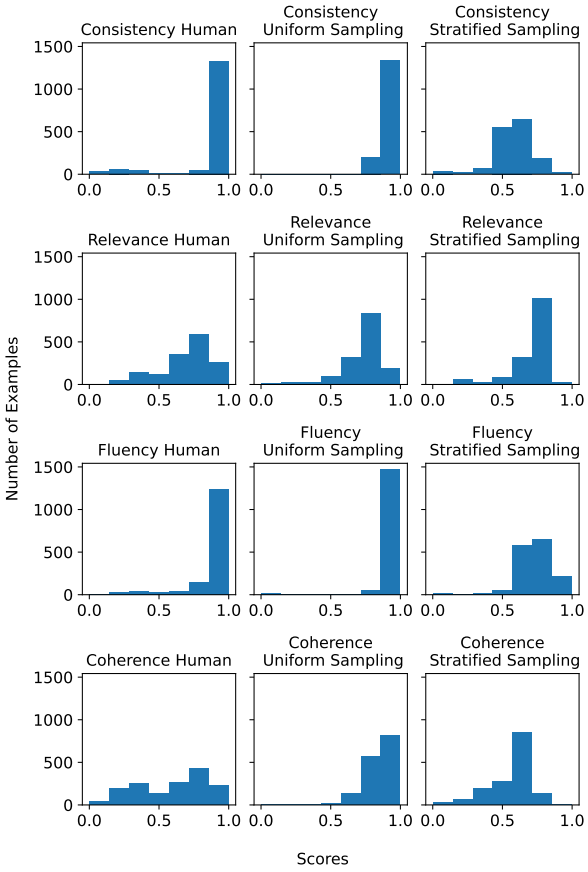

Figure 2: Distributions of human scores and predicted scores using ICE with uniform and stratified sampling on the SummEval benchmark

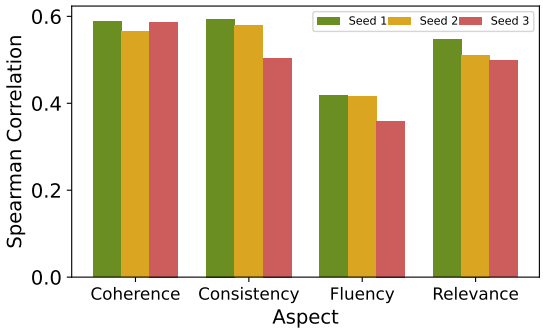

Figure 3: Effect of sampling different in-context examples. The performance over the same test set is observed to be robust to the choice of in-context examples.

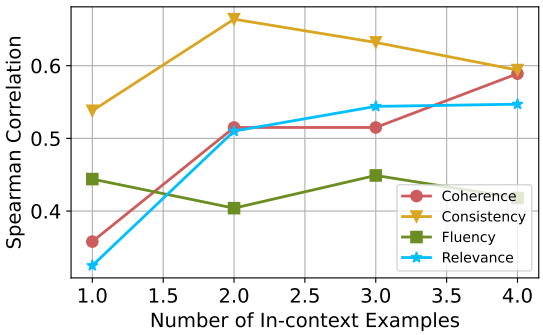

Figure 4: Effect of varying the number of in-context examples.

domain regardless of the true distribution. This forces predictions towards a centered distribution, which can cause the performance drop we observe in Table 1 when evaluating `consistency` using stratified sampling. Uniform sampling, on the other hand, selects examples that represent the true distribution, making model predictions more closely reflect the true distribution.

A drawback of uniform sampling is sub-optimal calibration in low-probability regions of the true distribution. For instance, if uniform sampling is used to evaluate `consistency`, the model might not see in-context examples with (say) scores less than 0.3 (Figure 2). This can affect output calibration in that region. Nonetheless, we suggest using uniform sampling in general. It is more stable and its prediction distribution closely follows the true distribution. For dimensions where it underperforms stratified sampling, the margins are less significant. Finally, even when ICE (uniform sampling) scores are calibrated differently from human scores, they still rank summary-quality correctly, insofar as our main results (Table 1) show that they compete with state-of-the-art on ranking-based metrics like Kendall-Tau and Spearman correlation. We use uniform sampling to select in-context examples in Sections 4.2 and 4.3.

4.2 Effect of Selection of In-context Examples

In order to determine whether performance is robust to the choice of in-context examples, we evaluate our test set using three different random sets of in-context examples. We observe in Figure 3 that for a given dimension, the maximum variation across three seeds is about 7 points, suggesting reasonably stable performance across the choice of in-context examples.

4.3 Effect of Number of In-context Examples

We evaluate our test set using different numbers of in-context examples (Figure 4). We observe that only for `relevance` and `coherence` does performance show improvement as we increase the number of examples. One reason for this could be the distribution of scores for a given dimension in the test set (Figure 2). Concretely, `consistency` and `fluency` mostly have near-perfect scores and therefore do not benefit from more samples while the

Metric	Model	Coh.	Con.	Flu.	Rel.	Overall
Human	GPT-3	4.85	4.73	4.97	4.65	4.80
	BRIO	4.57	4.65	4.88	4.48	4.65
	T0	4.15	4.47	4.78	3.68	4.27
ROUGE-L	GPT-3					22.09
	BRIO			-		28.20
	T0					26.63
BARTSc.	GPT-3	-1.25	-1.25	-1.25	-1.25	-1.25
	BRIO	-0.71	-0.71	-0.71	-0.71	-0.71
	T0	-0.96	-0.96	-0.96	-0.96	-0.96
ICE	GPT-3	0.908	0.996	0.994	0.849	0.937
	BRIO	0.896	0.993	0.993	0.834	0.929
	T0	0.890	0.981	0.985	0.761	0.904

Table 2: System-level scores from human annotations and automatic metrics. For each aspect, we color a given metric's highest/lowest rated system with orange/purple.

scores for `coherence` and `relevance` are spread out and therefore more samples allow representation over the whole range of scores.

Another observation is that even for `coherence` and `relevance`, performance with a single in-context example reaches near that achieved by some of the weaker fine-tuned baselines in Table 1. This suggests that the model possesses the notion of the evaluation task from pre-training itself, which is in line with recent work (Reynolds and McDonell, 2021; Min et al., 2022) that suggests that demonstrations help extract this knowledge.

Finally, we note that calibration can potentially be improved by increasing the number of examples. For instance, we observed that the four in-context examples that the uniform sampling procedure chose for coherence in Figure 2 had scores that fall between 0.7 and 1.0. This concentrates the prediction distribution in that range. The probability of such an event will reduce as the number of examples is increased further.

5 Using ICE to Evaluate Zero-Shot Prompting Models

Recent work by Goyal et al. (2022) showed that standard reference-based and reference-free metrics are not reliable in evaluating zero-shot summaries written by models such as GPT-3. Through a human study comparing summaries from three systems–GPT-3, BRIO, and T0–they observed that while humans prefer GPT-3 summaries, automatic evaluators consistently score GPT-3 summaries lower than summaries from other models.

We study the efficacy of ICE in evaluating zero-shot summaries written by GPT-3 at a dimension level. We use the set of 500 CNN articles from Goyal et al. (2022), with summaries from GPT-3, BRIO, and T0 for each article. We sample 100 of these articles and have three annotators rate summaries for each of the dimensions defined in Section 2.2 on a scale of $\{1, 2, 3, 4, 5\}$. We use ICE, ROUGE, and BARTScore (all of which do not require training data) to evaluate the summaries and present system-level results in Table 2.

We observe that ICE agrees with human judgments for each dimension and overall preferences while existing reference-based and reference-free metrics such as ROUGE and BARTScore[3] consistently rate GPT-3 summaries low. Goyal et al. (2022) suggest that most existing evaluation metrics reward summaries that imitate references, while GPT-3 summaries are zero-shot and not trained to imitate human-written references, which is likely why they are penalized by most existing evaluators. However, since ICE is not based on reference similarity (except when evaluating `relevance`) and is also not trained with reference summaries, it is able to better evaluate GPT-3 summaries and agrees with human preferences.

6 Conclusion

We show that in-context learning can be used for NLG evaluation as an alternative to fine-tuned evaluation metrics. Using a small number of examples, in-context learning evaluators can reach or exceed the state-of-the-art on multi-dimensional evaluation and that this is robust to the choice of in-context examples. Finally, we show that in-context learning evaluators align well with human judgements when evaluating summaries written by GPT-3.

Limitations

While ICE does not require fine-tuning on large amounts of data, it requires querying a powerful LLM at inference time (we use GPT-3 for our experiments which has 175 billion parameters). This can be a pay-per-use model or an open-source model such as BLOOM. This makes a downstream system that uses ICE reliant on an external dependency, which carries the risk of the external dependency failing.

[3]SummEval annotations are all based on the source, and the src-to-hyp version of BARTScore performs best across dimensions for this benchmark. We use this version for all dimensions, leading to identical scores. We format BARTScore results unlike ROUGE-L because in theory BARTScores can differ across dimensions for an arbitrary benchmark.

Relatedly, in this paper, we are limited due to monetary constraints in a variety of experiments we perform. For instance, we restrict ourselves to text summarization and use samples of benchmark meta-evaluation suites during some of our experiments. We leave the investigation of using ICE for other dimensions and downstream tasks for future work.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

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.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

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

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*.

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.

A Splitting SummEval and the Selection of In-context Examples

We randomly select 4 articles from the SummEval dataset and pick one system-generated summary from each article as an in-context example using the procedures outlined in Section 2.3. In other words, we pick $n = 4$ in Figure 1. For a given value of n, prompts for evaluating `consistency` are the longest since they contain entire source articles. We pick n such that `consistency` prompts fit within the context window of the model. We study the effect of the choice of n in Section 4.3.

To ensure that there is no overlap in the source article of any in-context example with the source article of any test example, we remove all summaries corresponding to the 4 selected source texts and use the remaining 1536 examples from SummEval as our test set. We ensure the absence of overlap throughout all experiments in Sections 3, 4, and 5.

B Sampling Procedures

B.0.1 Uniform Random Sampling

One summary is picked uniformly at random from the set of 16 summaries for a given source text. We do this for each of the 4 source texts selected to pick in-context examples from. Each of the 4 sampled in-context examples then consists of the selected summary, its human evaluation score on the current aspect of interest, and (optionally) the source text or the reference text.

B.0.2 Stratified Sampling

Let A denote the score of a summary on the aspect we are evaluating for; then $A \in [0, 1]$. In stratified sampling, we define 4 buckets by the ranges $\{[0, 0.25], (0.25, 0.5], (0.5, 0.75], (0.75, 1.0]\}$. We assign summary s to one of the buckets depending on the value of A_s. We do this for each of the 64 in-context example candidate summaries. Finally, we pick 4 summaries from the 64 candidates such that each summary falls into a different bucket and also comes from a different source text. We perform an exhaustive search for such an assignment, and in case no such assignment is possible (this can happen if none of the 64 summaries fall in a given bucket), we pick an arbitrary summary from a randomly selected bucket, ensuring that all 4 summaries come from different source articles.

For both uniform and random sampling, we ensure that each summary corresponds to a different source article.

C Annotation Procedure for Rating GPT-3, BRIO, and T0 Summaries

Summaries are annotated on a scale of $\{1, 2, 3, 4, 5\}$ for `coherence`, `consistency`, `fluency`, and `relevance` using the annotation instructions from Fabbri et al. (2020).

D Use of Existing Evaluation Packages

We use existing packages for all our baselines–ROUGE, BARTScore, CTC, and UniEval. For ROUGE, we use the native python implementation and report ROUGE-L scores for our experiment in Section 5. For BARTScore, we use the implementation accompanying the paper with the `source` to `hypothesis` setting across all dimensions, as that gives the best correlations with human judgments across dimensions. For UniEval, we use pre-trained model released by the authors to obtain results in Table 1 on the test set of size 1536.

E Significance Tests

Since ICE scores for some dimensions are close to UniEval scores, we perform pairwise tests to determine when one method is better than the other. Concretely, we compare performance on 1000 bootstrap samples by randomly selecting 80% of the test set for each sample. We observe that when using Kendall's Tau, ICE with uniform sampling outperforms UniEval with a significance level of 0.05 on both `consistency` and `relevance`. When using Spearman's rank correlation, Ice again outperforms UniEval on `consistency`, but the test is inconclusive at that significance level for `relevance`.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section at the end

☑ A2. Did you discuss any potential risks of your work?
In the limitations section, we discuss the risks associated with relying on external dependencies in deploying a framework such as the one studied in our work, if one intends to build a real-world application around it.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We generate ratings of system-generated summaries on the basis of quality

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. We created the artifacts

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. We have not, at the moment, decided on the term of distribution of our human-annotated data

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. The artifacts we create are built on top of publicly available datasets of publicly available news articles, which do not constitute data accessed solely for research purposes

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. We do not collect any new data. The data we use consists of CNN articles.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Our artifacts are ratings of system generated summaries from news articles. We mention this in the relevant section–Section 5.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
We mention these statistics for all our experiments in Sections 3, 4, and 5.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Almost all sections other than Introduction describe computational experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We mention in the paper that our backbone is GPT-3. We mention its number of parameters in the limitations section.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Most of the "hyperparamters" for our framework are prompt engineering choices, which we discuss extensively in Sections 4 and 5. We mention relevant parameters of our GPT backbone (such as sampling temperature) in Section 3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
A number of our analyses are done on samples of the benchmark datasets, and we have described where and how we are setting up and reporting multiple runs. We have added significance tests to validate results, where necessary.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix D

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 5

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. We use precisely the same instructions as used to annotate SummEval, our benchmark meta-evaluation dataset. We highlight the main points of the instructions in our paper but redirect readers to the original paper for the full text

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. We performed the relevant annotations

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. We annotate system-generated summaries of publically available news (CNN) articles for quality. We do not use/curate any individual's personal data.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. The source of the data is news articles. For our study, we annotate system-generated summaries of such articles for quality