ACUEVAL: Fine-grained Hallucination Evaluation and Correction for Abstractive Summarization

Anonymous ACL submission

Abstract

The impressive generation capabilities of large language models (LLMs) have made it even harder to detect the subtle hallucinations they make in abstractive summarization, where generated summaries consist of a blend of correct and incorrect information w.r.t. a given document. Recently-proposed LLM-based evaluation metrics attempt to capture this, but still face challenges: (1) they are biased towards summaries generated from the same underlying LLM, and (2) they lack interpretability, offering only a single score. In this work, we present ACUEVAL, a metric that leverages the power of LLMs to perform two sub-tasks: decomposing summaries into atomic content units (ACUs), and validating them against the source document. Compared to current strong LLM-based metrics, our two-step evaluation strategy improves correlation with human judgments of faithfulness on three summarization evaluation benchmarks by 3% in balanced accuracy compared to the next-best metric, and also shows reduced preference bias towards LLM-generated summary (by operating with fine-grained units). Further, we show that errors detected by ACUEVAL can be used to generate actionable feedback for refining the summary, successfully improving the faithfulness scores by more than 10%.[1]

1 Introduction

Hallucination in abstractive summarization, where the generation contains information that is inconsistent with the source document, remains a crucial problem despite the significant progress of large language models (LLM) (Goyal et al., 2022; Zhang et al., 2024). The problem has become more subtle, as the generations often contain a mixture of correct and hallucinated facts (Pagnoni et al., 2021; Min et al., 2023), making the detection of such errors harder. Recently-proposed evaluation metrics have achieved high correlations with human preferences with the aid of LLMs (Fu et al., 2023; Liu et al., 2023a). Nevertheless, similar to the observation by Tang et al. (2023) and Liu et al. (2023a), we find that such metrics generally have a *preference-bias*, where the metric favors generations from the same underlying LLM used for scoring. Furthermore, such metrics often output only a single numeric score, making them less interpretable to practitioners in understanding the precise location of the errors and the justification behind the score.

To address these problems, we present a new metric: ACUEVAL, which leverages the strong capability of LLMs to perform two fine-grained and structured sub-tasks instead of asking the model to directly provide a single score. We operate on the level of atomic content units (Liu et al., 2023b, ACUs), facts that can be verified and cannot be broken down further. ACUEVAL first generates these atomic facts from the system summary, and then validates each extracted fact against the source document. In Figure 1, we show that ACUEVAL successfully identifies that the second atomic fact is not consistent with the source document.

Operating on such fine-grained units as an intermediate representation instead of directly on the system summary reduces the preference bias of the metric in assigning high scores for summaries generated by the same underlying model. ACUEVAL involves two separate steps, each drawing on different input sources. The first step, ACU generation, relies solely on the system summary, while the second step, ACU verification, evaluates the consistency of the ACU with respect to the original document without the use of the summary. This separation ensures that the model does not implicitly assign the best score for the outputs generated by the same model. Moreover, the systematic matching between all extracted facts and the source document narrows down the issue of hallu-

[1] Our code will be made publicly available.

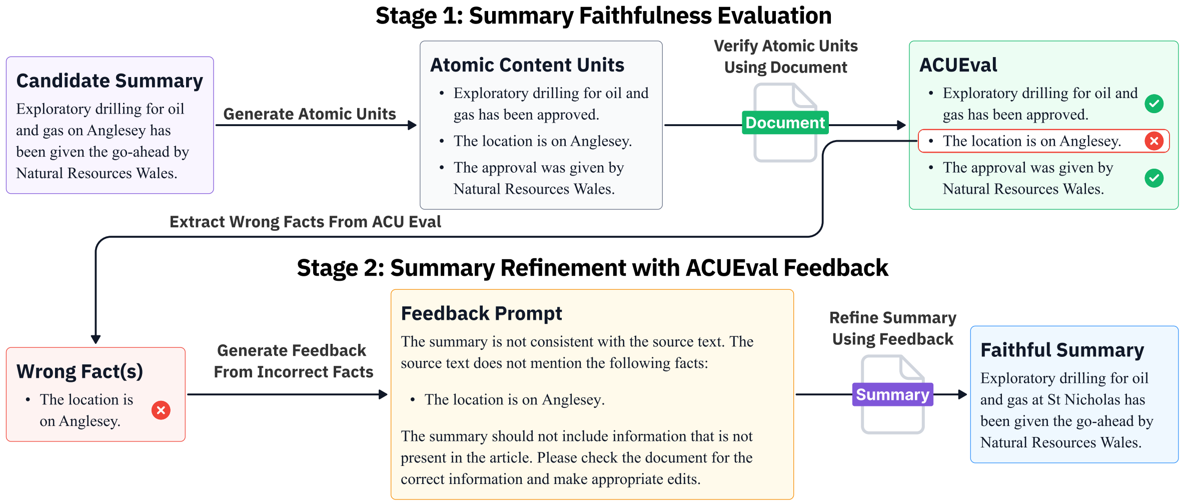

Figure 1: Illustration of ACUEVAL and its application in correcting hallucinations. For evaluation, the summary is broken down into atomic content units (ACUs), which are verified against the source document. For refinement, hallucinating ACUs are incorporated into the feedback prompt to improve the faithfulness of the summary.

cination precisely to the specific fact, allowing for better hallucination detection ability given the subtle mistakes LLMs make. The strong zero-shot and few-shot ability of LLMs also allow us to design a robust metric that can detect hallucinations across different datasets and system summaries without modifying the prompts for each setting.

First, we demonstrate that ACUEVAL aligns closely with human judgments across three summarization evaluation benchmarks (Fabbri et al., 2021; Tang et al., 2023; Zhang et al., 2024) and two datasets (Hermann et al., 2015; Narayan et al., 2018) which include summaries ranging from traditional approaches as well as those by recent powerful LLMs. ACUEVAL achieves higher correlations than previous metrics, including the recently proposed powerful LLM-based metrics. We show especially large improvements in detecting hallucinations for summaries generated by LLM-based models as opposed to summaries generated by traditional, fine-tuned models. Our detailed analysis in Section 5.2 also reveals that ACUEVAL significantly reduces the preference bias towards the summaries generated by the underlying LLM, due to operating on fine-grained units, unlike metrics that directly evaluate on the generated summary.

A novel downstream application of ACUEVAL's fine-grained error localization is to create detailed, structured feedback to improve faithfulness in the iterative summarization process (Zhang et al., 2023), where a refinement model addresses the problems listed in the comment to produce an enhanced summary. As shown at the bottom of Figure 1, all facts judged to be incorrect by ACUEVAL are incorporated into the feedback. By covering the detailed hallucinations detected by ACUEVAL, as demonstrated in Section 5.4, the targeted feedback informed by ACUEVAL enhances the model's ability to generate more faithful summaries after revision, leading to a 10% and 23% improvement on G-Eval (Liu et al., 2023a) and ACUEVAL scores compared to using GPT-4's feedback.

Finally, we provide an analysis using ACUEVAL to assess the capacity of various LLM's to produce faithful summaries. We first confirm ACUEVAL's effectiveness in identifying patterns consistent with those noted according to human annotations, specifically on the news summarization meta-evaluation benchmark (Zhang et al., 2024). Our findings, specifically that instructions-based models perform better and the reference summary achieves low faithfulness scores, align closely with human judgments. Next, we apply ACUEVAL to assess LLMs in the hallucination benchmark,[2] and find that GPT4 exhibits the least hallucination among the tested models, in line with previous findings (Min et al., 2023; Laban et al., 2023).

In summary, our contributions are the following:

1. We introduce ACUEVAL, an interpretable, LLM-based faithfulness evaluation metric for summarization, with a structured, two-step evaluation strategy that first breaks the output into fine-grained ACUs and then verifies their

[2] https://github.com/vectara/hallucination-leaderboard

presence in the source document.

2. We show that ACUEVAL achieves a higher correlation to human judgments than current LLM-based metrics, especially for LLM-generated summaries. With ACUEVAL, we observe trends such as GPT-4 containing the least hallucinations when assessing LLMs' capability to generate faithful summaries.
3. We show that the hallucinating ACUs detected by ACUEVAL can be in turn transformed into detailed actionable feedback for refining the summary for improved faithfulness.

2 Related Work

Faithfulness evaluation for summarization. Numerous metrics have been designed to assess the faithfulness of abstractive summarization. These range from entailment-based metrics (Kryscinski et al., 2020; Goyal and Durrett, 2020), to question-generation, question-answering metrics (Durmus et al., 2020; Scialom et al., 2021; Fabbri et al., 2022). More recently, the focus has shifted towards LLM-based metrics (Liu et al., 2023a; Fu et al., 2023; Gao et al., 2023; Luo et al., 2023) that leverage LLMs to assess the faithfulness of a summary. Our method uses an open-source LLM and splits the evaluation into two distinct sub-tasks, enhancing interpretability and mitigating the bias inherent to using the same LLM for both generation and evaluation.

Fine-grained metrics. The adoption of fine-grained units in summarization evaluation has improved inter-annotator agreement for human evaluation (Liu et al., 2023b; Krishna et al., 2023) as well as metric performance for automatic evaluation. For instance, DAE (Goyal and Durrett, 2020) outperforms traditional entailment-based metrics by focusing on the entailments of dependency arcs. Similarly, fine-grained units are also effective for relevance, dating back to works including Nenkova and Passonneau (2004); Shapira et al. (2019) that generate ACUs from the reference summary and validates against the system summaries. FactScore (Min et al., 2023) also uses a two-step approach for evaluating the factuality of people biographies. ACUEVAL, while sharing similarities with FactScore, leverages a single, open-sourced LLM for both generating and evaluating atomic facts, allowing for cost-effective and easily replicable future developments. For abstractive summarization, we do not need a separate retriever to source relevant passages, given that the source document is already provided. We also demonstrate that the fine-grained metrics can be useful beyond evaluation, improving the downstream summarization performance when used as feedback for refining the summary.

3 ACUEVAL

Figure 1 illustrates our metric ACUEVAL. Here, we assume that we have a source document X and a generated summary $\hat{y}$. ACUEVAL consists of two structured steps: (1) deconstructing the summary into fine-grained ACUs, and (2) predicting the presence of each ACU against the information presented in the source document. The result of these steps is a faithfulness score.

ACU generation. We first generate atomic content units (ACUs), or atomic facts, from the summary. We follow the definition of ACUs by Liu et al. (2023b): *Elementary information units, which no longer need to be further split for the purpose of reducing ambiguity in human evaluation.* We note that, unlike previous approaches where atomic facts were generated from reference summary y, we apply this method to the generated summary $\hat{y}$. This approach yields more fine-grained information of the summary, which has been shown to improve faithfulness evaluation (Goyal and Durrett, 2020; Durmus et al., 2020). Additionally, we opt for a textual representation over complex representations like dependency parses (Goyal and Durrett, 2020) or AMR graphs (Ribeiro et al., 2022), which simplifies the integration of error localization in LLMs. Formally, we break down a summary $\hat{y}$ into a list of atomic facts $A_{\hat{y}} = \{a1, a2, ..., a_N\}$. We generate these facts by asking an LLM to break an utterance up using the prompt shown in Figure 3.

ACU verification. After generating the ACUs, we then verify whether they are consistent with the source document X. This is done by prompting an LLM to predict whether each fact is consistent with the information in the source document with either "Yes" or "No" as the answer (See Figure 4). To refine our accuracy, we normalize the probability of the two labels and take the probability for "Yes" as the final score of the ACU. We use same LLM for both ACU generation and ACU verification. Formally, the score for each ACU is defined as:

$$s_i = p(\text{LLM}(X, a_i, pt) = \text{Yes})$$

where $\text{LLM}(X, a, pt)$ is LLM's prediction given the document X, the ACU a, and the prompt pt.

Final Score. The final score is the average across all ACU presence predictions:

$$\text{ACUE}\textsc{val} = \frac{1}{|A|}\sum_{i=1}^{|A|} s_i$$

Fine-grained Feedback from ACUEVAL. Next, we also demonstrate a novel application of fine-grained error localization with ACUEVAL: Generating detailed feedback based on the hallucinations identified by ACUEVAL for improving the summary. Inspired by Saunders et al. (2022), who demonstrated that model-generated critiques can guide humans to detect overlooked flaws, our approach similarly uses detailed critiques to assist the refinement model in identifying and correcting hallucination. The refinement process with ACUEVAL can be seen in the lower section of Figure 1. Unlike the original method where the critique model generates free-form feedback, our strategy involves listing all ACUs deemed inconsistent with the document as inconsistent facts that the refinement model needs to address (see Figure 6 for the prompt template). Since the original critique model itself is quite similar to the LLM-based metrics proposed in prior works,[3] it might overlook certain hallucinations because of the model's coarse-grained scope and inherent preference biases. In contrast, the advantage that ACUEVAL has over critique models when used for feedback mirrors its benefits for evaluation purposes, where ACUEVAL offers a more exhaustive detection of hallucinations with little preference bias.

4 Experiments

4.1 Implementation Details

We use StableBeluga 2 (Mahan et al., 2023) for both ACU generation and ACU verification, as we find that this model follows the instruction reliably.[4] The model uses Llama2 70B (Touvron et al., 2023) as the backbone, and fine-tuned on the ORCA (Mukherjee et al., 2023) dataset. We use greedy decoding to ensure determinism and set the maximum generation length to 256 for ACU generation and 5 for ACU verification. More details can be found in Appendix B.

4.2 Benchmarks

We focus on abstractive summarization benchmarks that measure summary faithfulness by collecting human judgments. All of the benchmarks consist of examples from two news summarization datasets CNN/DM (Hermann et al., 2015) and XSum (Narayan et al., 2018), containing news articles from CNN/Dailymail and BBC, respectively. We include benchmarks consisting of annotations on summaries generated by previous state-of-the-art models, as well as those by recent LLM models. **SUMMEVAL** (Fabbri et al., 2021) consists of annotations from extractive and abstractive systems. **AGGREFACT** (Tang et al., 2023) consists of 9 faithfulness benchmark datasets. We use the FTSOTA split consisting of state-of-the-art fine-tuned summarization models, as the authors find that previous metrics, including LLM-based metrics, fall short when evaluating summaries from more recent models. **LLMSUMMEVAL** (Zhang et al., 2024) is our primary evaluation benchmark, consisting of similar human annotations on summaries generated by LLMs under both zero-shot and few-shot settings. More details can be found in Appendix C.

4.3 Evaluation

Given the issue of significant class imbalance in the data, computing correlations directly to human labels may not accurately reflect performance. This problem is particularly crucial in contexts like the LLMSUMMEVAL benchmark, where only 20% of annotations are marked as incorrect. To mitigate the impact of this imbalance, we follow Laban et al. (2022); Tang et al. (2023) and focus on computing balanced accuracy. To ensure a fair evaluation across the diverse scales of metric scores, we additionally split the annotations into validation and test sets based on whether their indices are odd or even, following Tang et al. (2023). This allows us to tune for the threshold for the optimal balanced accuracy within the validation dataset. Experiments on standard correlations are in Appendix E.

4.4 Baseline Metrics

We include baseline metrics in the respective benchmarks as well as strong faithfulness metrics developed for summarization evaluation. Our pri-

[3] Both the critique model and LLM-based metrics, such as G-Eval, take the document and summary as input and output a text. However, while LLM-based metrics generate a score reflecting the quality of the summary, the critique model produces a textual commentary of the summary's content.

[4] We have also tried Llama2-chat, Zhepyr, and Vicuna 33B, but we find that they do not follow the prompt consistently (i.e. only predicting numerical scores or only answering true/false).

	SUMMEVAL	AGGREFACT-FTSOTA		LLMSUMMEVAL	
		CNN/DM	XSum	CNN/DM	XSum
DAE	64.8 ± 2.4	65.4 ± 4.4	70.2 ± 2.3	84.6 ± 1.7	72.9 ± 1.6
QuestEval	73.8 ± 2.6	70.2 ± 3.2	59.5 ± 2.7	86.5 ± 1.9	75.1 ± 1.5
QAFactEval	83.0 ± 1.7	67.8 ± 4.1	63.9 ± 2.4	68.3 ± 3.4	62.3 ± 2.0
ChatGPT-ZS	-	56.3 ± 2.9	62.7 ± 1.7	-	-
G-Eval	81.9 ± 1.5	-	-	-	-
BelugaEval	81.1 ± 1.6	56.1 ± 2.8	66.1 ± 1.7	77.0 ± 2.0	62.8 ± 1.7
ACUEVAL	**86.2** ± **2.1**	**70.4** ± **3.3**	**74.5** ± **1.7**	**89.5** ± **1.6**	**78.4** ± **1.5**

Table 1: Balanced accuracy on summarization benchmarks with 95% confidence intervals.

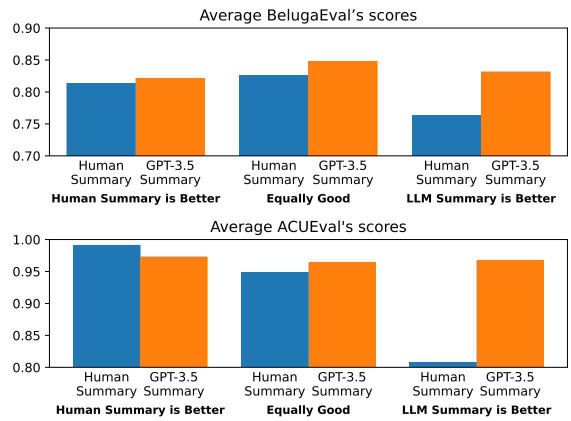

Figure 2: Average metric scores for human written summaries and GPT-3.5 summaries. We convert BelugaEval's score to the same scale as ACUEVAL. ACUEVAL (bottom) closely aligns its scoring with human judgments, awarding higher scores to human summaries deemed superior by annotators, and lower scores when the opposite is true. In contrast, BelugaEval (top) reveals a preference bias towards GPT-3.5 summaries by assigning them higher scores across all settings.

mary focus is a comparison with LLM-based metrics, which have been shown to be better at detecting hallucination than traditional metrics. We include strong GPT-based metrics, including G-Eval (Liu et al., 2023a), ChatGPT-ZS (Luo et al., 2023). However, due to the high cost of running the metric across all benchmarks, we also explore an alternative, BelugaEval, our variant of G-Eval and ChatGPT-ZS based on StableBeluga 2. This open-source model offers a similar approach and performance to G-Eval. Finally, we also include standard faithfulness metrics, including DAE (Goyal and Durrett, 2020), QuestEval (Scialom et al., 2021), and QAFactEval (Fabbri et al., 2022). For more information, we refer the readers to Appendix A.

5 Results

5.1 Meta-Evaluation

We present the balanced accuracy results on the three benchmarks in Table 1. We first note that BelugaEval is a reliable alternative to G-Eval and ChatGPT-ZS, as it achieves similar balanced accuracy that differs at most by 1 point. For XSum split of AGGREFACT-FTSOTA, BelugaEval improves 3.4 points over ChatGPT-ZS. ACUEVAL consistently achieves the highest balanced accuracy on all three evaluation benchmarks. Notably, in LLM-SUMMEVAL, our main benchmark, ACUEVAL surpasses the next-best metric by 3 points in both CNN/DM and XSum datasets, highlighting the high accuracy and robustness of ACUEVAL.

Interestingly, despite showing high correlations with human judgments, the LLM-based evaluation metrics, including G-Eval, ChatGPT-ZS, and BelugaEval, do not outperform some of the more established baseline metrics in terms of balanced accuracy. Particularly, the LLM-based metrics' performance was the lowest for the CNN/DM split of AGGREFACT-FTSOTA. This aligns with findings from Tang et al. (2023), which suggest that while these metrics excel in assessing outputs from older systems, they may not be as effective with content generated by more recent models.

Furthermore, the results reveal that different metrics show varied trends when assessing summaries produced by earlier systems compared to those generated by LLMs. For instance, while QuestEval had the lowest balanced accuracy for AGGREFACT-FTSOTA XSum benchmark, it achieved the highest accuracy among traditional metrics in the LLM-SUMMEVAL XSum benchmark. This underscores the importance of re-evaluating various metrics, especially in the context of LLM-generated content, which differs from traditional benchmarks.

	CNN/DM	XSum	Samsum
LLM FactScore	69.33	**72.55**	76.86
LLM 1-shot	71.90	66.88	75.44
LLM 3-shot	76.48	71.21	81.28
LLM 5-shot	**76.59**	70.86	**81.36**
AutoACU2-gen	84.07	82.00	86.96

Table 2: ACU generation results of different prompts on the ROSE dataset.

	CNN/DM	XSum	Samsum	Average
DeBERTa-XLarge	60.21	**73.09**	62.77	65.36
LLM FactScore	61.71	52.35	63.80	59.29
LLM Zeroshot	**78.32**	70.93	**82.00**	**77.08**
LLM Fewshot	77.69	71.52	81.86	77.02
LLM Fewshot + doc.	40.50	28.85	40.85	36.73
AutoACU2-match	91.58	92.85	90.80	91.75

Table 3: ACU verification results of different prompts on the ROSE dataset.

5.2 Preference over LLM-based Outputs

A key concern with metrics based on LLMs is their potential bias towards outputs generated by similar LLMs. This issue arises because these metrics often use the same or a related generation model for evaluation, leading to higher scores for outputs from similar models (Deutsch et al., 2022).

Liu et al. (2023a) observed a tendency for G-Eval to favor outputs from GPT-3.5 models over human-written summaries. To investigate this, we conducted similar experiments comparing the metric scores for human-written summaries with those generated by GPT-3.5. We split the GPT-3.5 summaries from LLMSUMMEVAL into three categories based on how they were rated against human summaries: higher, equal, or lower, and compare the average metric scores for human summaries and the GPT-3.5 summaries under the three cases.

We perform the analysis using BelugaEval and ACUEVAL, both of which use the same underlying LLM and present the result in Figure 2. We see a clear bias in BelugaEval: It often rates GPT-3.5 summaries higher than human-written ones, even when human annotators preferred the latter. In the figure, we see that the average BelugaEval scores of GPT-3.5 summary are always higher than that for the human summaries. However, our metric, ACUEVAL, demonstrated more balanced behavior, assigning higher scores to superior human summaries and vice versa. Nevertheless, it still shows a slight preference for GPT-3.5 summaries where the summaries were deemed equally good.

5.3 Ablations

ACU generation performance. We wish to evaluate the effectiveness of the ACU generation capability. Since there are no gold ACUs available for generated summaries for comparison, we use the ROSE dataset (Liu et al., 2023b) which provides expert-written ACUs for examples in the CNN/DM, XSum, and Samsum dataset (Gliwa et al., 2019). We compare the ACUs generated by ACUEVAL with expert-written ones. Following the authors, we calculate the Rouge1-F1 (Lin, 2004) score for each generated ACU by greedily finding its best match among the reference ACUs and then taking the average across all generated ACUs.

We experiment with various prompts, including those from FactScore (Min et al., 2023), and create few-shot prompts using the gold ACUs from ROSE's CNN/DM validation set. Details on our prompt design can be found in the Appendix G.1. We include AutoACU2-gen (Liu et al., 2023b), a T0-3B (Sanh et al., 2022) model fine-tuned on all the reference ACUs as a potential upper bound for this task. Table 2 show the impact of different prompt strategies on ACU quality.

We observe that providing more context-specific examples (from 3-shot to 5-shot) leads to marginal improvements for the CNN/DM and Samsum datasets. The FactScore prompt, which focuses on sentence-level generation, shows better results for XSum, which contains single-sentence summaries. However, since expert-written ACUs are typically based on multi-sentence summaries and include cross-sentence references, the FactScore approach falls short for CNN/DM and Samsum, often resorting to generic subject assignments.

In conclusion, though there still exists a large gap between the few-shot approach and the full fine-tuning method, our analysis indicates that the 5-shot variant is preferable. It not only achieves the highest Rouge1 score for the CNN/DM dataset among the different prompts but also demonstrates robust performance across different types of summaries, including those in the Samsum dataset.

ACU verification performance. Next, we evaluate the ACU verification capability. We again utilize the ROSE dataset containing expert labels for the presence of reference ACUs in candidate summaries and evaluated using the accuracy of as-

	CNN/DM		XSum		All	
	G-Eval	ACUEval	G-Eval	ACUEval	G-Eval	ACUEval
Original Summaries	3.11	50.7	2.50	57.9	2.80	54.3
G-Eval Feedback	4.53	75.7	4.34	82.6	4.43	79.1
ACUEval Feedback	**4.97**	**97.3**	**4.79**	**97.0**	**4.88**	**97.1**

Table 4: Faithfulness scores on refined summaries with different feedbacks.

signing the correct label. It is important to note that this task is slightly different from the standard usage of the ACUEval: here, we generate ACUs from a reference summary (y) and compare them to a candidate summary ($\hat{y}$). In contrast, for evaluating faithfulness, ACUs are derived from the candidate summary ($\hat{y}$) and then matched against the original document (X).

Similar to the ablations on ACU generation capability, we explore different prompts, including FactScore-style prompts, as well as zero-shot and few-shot approaches. We also investigate the impact of incorporating the document as additional input in a few-shot prompt setup. To compare to other metrics for this subtask, we include AutoACU2-match, a model using DeBERTa-XLarge, trained on all the ACU-summary pairs in the ROSE dataset, alongside the original pre-trained model.

We present our results in Table 3. Notably, the zero-shot prompt technique emerges as the most accurate, surpassing the results of the pre-trained DeBERTa-XLarge model by 12 points across the three datasets, and is slightly better than the few-shot variant. The FactScore prompt here does not show adaptability to our model and task, as it does not achieve high accuracy. Adding the document to the prompt also results in a noticeable decrease in performance. This observation is consistent with the findings of Liu et al. (2023c), who noted that inputs with redundant information could negatively impact predictive performance.

5.4 Improving Generation via Feedback

In this section, we assess the impact of the feedback informed by ACUEval on improving summary faithfulness in the summary refinement pipeline discussed in Section 3. ACUEval feedback consists of a comprehensive list of the atomic facts that are judged as incorrect according to ACUEval. We compare our feedback generation method against the self-critique method, where GPT-4 is tasked to provide a critique of the summary directly. This is achieved by asking the model to continue producing content after it has assigned the faithfulness score with the G-Eval prompt.

For a fair comparison, we use GPT-4 as the refinement model with the same refinement prompt but change the feedback depending on the method. Examples of the refinement prompt with ACUEval feedback and G-Eval feedback can be seen in Figure 6 and Figure 7, respectively. The refinement model takes as input a prompt consisting of the document, summary, and feedback. Note that the setup is similar to the iterative summarization process (Zhang et al., 2023), but we also include the original document as additional input. This is important for comparing the two methods fairly because ACUEval only highlights where the summary deviates from the source document but does not provide the correct content directly.

We randomly selected 50 summaries each from CNN/DM and XSum within the LLMSummEval dataset, all containing errors identified by ACUEval. We measure the faithfulness of the refined summary using ACUEval and G-Eval, which allows us to verify that the gain does not stem from optimizing on our proposed ACUEval. The result in Table 4 shows that both feedback types improve summary faithfulness, but ACUEval feedback leads to the most substantial improvement. Summaries refined with ACUEval feedback nearly reached perfect faithfulness score, achieving 4.88 out of 5 for G-Eval and 97.1 for ACUEval. This highlights the strength of ACUEval at providing nuanced feedback, enabling LLMs to significantly improve on faithfulness.

6 Benchmarking LLM for Faithfulness

Lastly, ACUEval can also serve as a powerful analytical tool for assessing the capacity of current Large Language Models (LLMs) to generate faithful summaries. Demonstrating strong correlations with human evaluations, particularly in recent models, ACUEval provides a practical and reliable alternative to human assessments of hallucinations.

6.1 LLMSummEval

We first examine the ACUEval scores of various models using LLMSummEval in Table 6, which

Model	ACUEVAL	HEM	Summary Length (# words)	Answer Rate (%)
GPT-4	0.995	0.970	81.1	100.0
GPT-3.5	0.992	0.965	84.1	99.6
Llama 2 (70B)	0.989	0.949	84.9	99.9
Anthropic Claude 2	0.988	0.915	87.5	99.3
Llama 2 (13B)	0.984	0.941	82.1	99.8
Google Palm (text-bison-001)	0.977	0.879	36.2	92.4
Cohere (52.4B)	0.970	0.915	59.8	99.8
Cohere-Chat (52.4B)	0.967	0.925	74.4	98.0
Llama 2 (7B)	0.966	0.944	119.9	99.6
Mistral (7B)	0.962	0.906	96.1	97.6
Google Palm-Chat (chat-bison-001)	0.755	0.728	221.1	88.8

Table 5: Hallucination Benchmark sorted by ACUEVAL. We include the model size when possible.

allows us to compare against the provided human judgments. The high congruence of these scores with human ratings indicates our metric's alignment with human judgment. Our findings echo the insights Zhang et al. (2024) in several ways: we find that *Instruction-tuned models perform better*, and *reference summaries are less faithful*. More detailed discussions can be found in Section D. In summary, ACUEVAL's scoring closely aligns with human judgments, demonstrating its efficacy as a benchmarking tool for discovering informative trends among the models.

6.2 Hallucination Benchmark

To compare the faithfulness power of more recent popular LLMs, we also calculate the ACUEVAL scores on the hallucination benchmark. It contains summaries of 831 documents using 11 strong LLMs. The result is presented in Table 5. Models are ranked based on their performance according to ACUEVAL. The benchmark originally uses the Hallucination Evaluation Model[5] (HEM) as the benchmarking metric, which is trained on fact verification with DeBERTaV3 (He et al., 2023). Our ACUEVAL ranking reveals similar trends as observed with HEM: Models that maintain an optimal answer rate and adhere to average summary lengths tend to score higher in faithfulness.

In line with previous works (Min et al., 2023; Laban et al., 2023), GPT-4 and GPT-3.5 achieve the highest faithfulness scores among the models. Apart from these two models, there is no single model family that consistently shows improvement with scaling model size on HEM-based ranking. Nevertheless, our ACUEVAL-based ranking reveals a notable phenomenon - *faithfulness scales with model size within the same model family*. Our ranking underscores a clear correlation between model size and faithfulness in producing summaries of comparable lengths. For instance, the Llama 2 series shows a definitive hierarchy in faithfulness: 70B outperforms 13B, which in turn surpasses 7B. Similarly, Mistal 7B aligns closely with Llama 2 7B in terms of ranking. This contrasts with the HEM ranking, where a distinct hierarchy is evident (GPT > Llama 2 > Cohere > Claude 2 > Mistral > Google Palm).

[5] https://huggingface.co/vectara/hallucination_evaluation_model

7 Conclusion

In this paper, we introduce ACUEVAL an interpretable, fine-grained metric for evaluating faithfulness for abstractive summarization. Our findings demonstrate that ACUEVAL achieves the highest balanced accuracies across diverse benchmarks and datasets, outperforming other recent, strong LLM-based metrics. Notably, ACUEVAL shows very low bias towards LLM-generated outputs, making it a fair tool for evaluation of summaries in the era of LLMs. Next, we also explore how ACUs that are considered not faithful to the input document can be incorporated as detailed feedback, which in turn enhances the correction model at refining the summary with little hallucination. Finally, we compare the average ACUEVAL scores of various LLMs, assessing their faithfulness in abstractive summarization. These comparisons align closely with human judgments and reveal that larger models tend to be more faithful. We hope that ACUEVAL can serve as a foundational guide for evaluating generated summaries. Looking forward, we propose expanding this framework to encompass additional facets of summarization evaluation and adapting it for more complex tasks like multi-document and long-form summarization.

Limitations

One key challenge of our current approach is the slow computation stemming from the need to break down the evaluation into two sub-tasks. This issue becomes particularly evident as the length of the summary grows, resulting in an increased number of small elements that must be individually checked. A potential solution is to have the model verify a variable number of ACUs in a single step, rather than one at a time. Another possibility is to merge two separate steps: having the model both create and then verify these elements in one go. However, this may be not reliable, as current LLMs cannot accurately follow multiple steps at once. Another limitation is the need for a model that can accurately follow instructions. We tested various LLMs, but many struggled with either generating or verifying the ACUs accurately. Mistakes in the generation phase can lead to further errors down the line, magnifying the problem, which is true for all model-based evaluation methods.

References

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

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

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

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. 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.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. 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.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Alexander Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. Towards interpretable and efficient automatic reference-based summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16360–16368, Singapore. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Dakota Mahan, Ryan Carlow, Louis Castricato, Nathan Cooper, and Christian Laforte. 2023. Stable beluga models.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of* the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. 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 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. SummIt: Iterative text summarization via ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657, Singapore. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

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

A Details on Baseline Metrics

G-Eval (Liu et al., 2023a) is a GPT4-based metric, which significantly outperforms traditional metrics. However, due to the high cost of running the metric across all benchmarks, we also explore an alternative, BelugaEval. This open-source model offers a similar approach and performance. See Appendix F for more details on the comparison between the two metrics.

ChatGPT-ZS (Luo et al., 2023) is another LLM-based metric that uses ChatGPT to evaluate summaries. We include the results included by AGGREFACT. Similarly, BelugaEval can be seen as an alternative metric.

BelugaEval is our variant of G-Eval and ChatGPT-ZS based on StableBeluga 2. We use a similar prompt as G-Eval, which can be found in Figure 5. Following Liu et al. (2023a), we integrate a chain-of-thought prompt and utilize the score normalization technique, where the final score is calculated as the weighted summation of the 1-5 scale, each weighted by its respective normalized probability. We refer the readers to the original paper for more details.

DAE (Goyal and Durrett, 2020) is a fine-grained entailment metric that evaluates the faithfulness between the summary's dependency arcs and the document.

QuestEval (Scialom et al., 2021) is a question-generation question-answering (QGQA) metric. It computes answer overlap scores by generating questions from a source document and then assessing how well these questions are answered by the summary, and vice-versa.

QAFactEval (Fabbri et al., 2022) is a highly optimized QGQA-based metric after extensive analysis of the individual components.

B Implementation Details

We use 8 A100s GPUs to run all experiments. Running the ACU generation for each benchmark takes around 8 hours and running the ACU verification takes around 10 hours. Running GPT-4 (*gpt-4-0613*) for refinement takes around 10 minutes for the 50 examples. For all of our experiments, we use the transformers package (Wolf et al., 2020). All baseline metrics are used with the corresponding official implementations. For calculating balanced accuracy and correlations, we use the official scripts from AggreFact (Tang et al., 2023) and ROSE dataset (Liu et al., 2023b), respectively. Since we use greedy decoding for all experiments for deterministic behavior, we only perform single runs for all experiments.

C Benchmark Details

SUMMEVAL (Fabbri et al., 2021) consists of expert annotations of 100 samples from 17 different extractive and abstractive systems, all using the CNN/DM dataset (Hermann et al., 2015). To have a fair comparison to previous metrics, we use the first 16 systems that were part of the initial release. We use the *consistency* labels for assessing faithfulness. The labels are on a 1-5 Likert scale, and we convert the scores into binary labels following Laban et al. (2022): If the majority of the expert annotators award a summary a score of 5, the summary is categorized with 1.

AGGREFACT (Tang et al., 2023) consists of 9 faithfulness benchmark datasets on both CNN/DM and XSum (Narayan et al., 2018). This benchmark splits the summaries systems into three categories: FTSOTA, EXF, and OLD, representing state-of-the-art fine-tuned summarization models, early transformer models, and older models, respectively. All annotations are transformed to a binary label. We refer the readers to the original paper for more details. The authors find that previous metrics, including LLM-based metrics, tend to show high accuracy with older summaries but fall short when evaluating summaries from more recent models. We thus focus on the FTSOTA split, containing outputs generated by BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and PEGASUS (Zhang et al., 2020). The annotations are split according to the two datasets.

LLMSUMMEVAL (Zhang et al., 2024) is our primary evaluation benchmark. It collects similar human annotations on summaries generated by LLMs under both zero-shot and few-shot settings. It includes 100 examples of 19 settings for both CNN/DM and XSum.[6] Similar to AGGREFACT, we split the benchmark based on the two datasets.

All benchmarks use CNN/DM and XSum, which are under the MIT license. For the benchmarks, SummEval is under the MIT. We follow the authors' instructions for download and usage.

D Benchmarking on LLMSUMMEVAL

We first examine the ACUEVAL scores of various models using LLMSUMMEVAL in Table 6, which allows us to compare against the provided human judgments. The high congruence of these scores with human ratings indicates our metric's alignment with human judgment. Our findings echo the insights Zhang et al. (2024) in several ways:

Instruction-tuned models perform better. Instruction-tuned GPT-3 models, especially in zero-shot scenarios, surpass their non-instruction-tuned counterparts and generally achieve the highest faithfulness scores across datasets. Similar observations can be made under the few-shot setting for XSum. This trend also manifests under ACUEVAL scores, which show higher scores for instruction-tuned models. The only exception is the few-shot 350M model on CNN/DM, where human scores also consider the non-instruction-tuned models to be better.

Reference summaries are less faithful. Zhang et al. (2024) note that the reference summaries are poor for the two datasets. This can be directly verified, as the human scores for the reference summary are generally among the lowest ones, especially for XSum. ACUEVAL scores mirror this trend, placing reference summaries among the lowest.

In summary, ACUEVAL's scoring closely aligns with human judgments, demonstrating its efficacy as a benchmarking tool for discovering informative trends among the models.

[6] For XSum, the authors note that the 350M GPT3 model provides only empty outputs, and thus the XSum annotations

Setting	Models	CNN/DM ACUEVAL	CNN/DM Human	XSum ACUEVAL	XSum Human
Zero-shot	GPT-3 (350M)	0.287	0.29	0.277	0.26
	GPT-3 (6.7B)	0.267	0.29	0.688	0.77
	GPT-3 (175B)	0.511	0.76	0.416	0.80
	Ada Instruct v1 (350M)	0.817	0.88	0.878	0.81
	Curie Instruct v1 (6.7B)	0.986	0.97	**0.966**	0.96
	Davinci Instruct v2 (175B)	**0.992**	**0.99**	0.944	**0.97**
Few-shot	Anthropic-LM (52B)	**0.995**	0.94	0.926	0.70
	Cohere XL (52.4B)	0.962	**0.99**	0.883	0.63
	GLM (130B)	0.974	0.94	0.896	0.74
	OPT (175B)	0.989	0.96	0.891	0.67
	GPT-3 (350M)	0.891	0.86	-	-
	GPT-3 (6.7B)	0.960	0.97	0.864	0.75
	GPT-3 (175B)	0.991	**0.99**	0.858	0.69
	Ada Instruct v1 (350M)	0.817	0.84	0.736	0.63
	Curie Instruct v1 (6.7B)	0.988	0.96	0.928	**0.85**
	Davinci Instruct v2 (175B)	0.994	0.98	**0.940**	0.77
Fine-tuned	BRIO	0.983	0.94	**0.845**	**0.58**
	PEGASUS	**0.990**	**0.97**	0.842	0.57
References		0.968	0.84	0.785	0.37

Table 6: ACUEVAL scores on LLMSUMMEVAL benchmark.

	SUMMEVAL Sys.	SUMMEVAL Sum.	LLMSUMMEVAL-CNN Sys.	LLMSUMMEVAL-CNN Sum.	LLMSUMMEVAL-XSUM Sys.	LLMSUMMEVAL-XSUM Sum.
QuestEval	0.700	0.271	0.578	0.406	0.556	0.423
UniEval	**0.750**	0.356	**0.637**	0.353	0.346	0.348
G-Eval	0.600	**0.463**	-	-	-	-
BelugaEval	0.700	0.403	0.472	0.364	0.425	0.270
ACUEVAL	0.683	0.369	**0.637**	**0.409**	**0.556**	**0.439**

Table 7: Kendall Correlation on SUMMEVAL and LLM-SUMMEVAL for *Consistency*.

E Results on Meta-Evaluation

Table 7 shows results with traditional meta-evaluation metrics, i.e. Kendall correlations on SUMMEVAL and LLMSUMMEVAL. The correlation results mirror the results we previously observed in balanced accuracy shown in Table 1. Notably, BelugaEval, representing LLM-based approaches that generate direct scores, shows a weaker correlation for more recent outputs from LLMSUMMEVAL. These correlations generally fall below those of baseline metrics. However, ACUEVAL achieves the highest system-level and summary-level correlations on both LLMSUMMEVAL benchmarks, especially on the XSum dataset, corresponding to the larger presence of hallucinations in the XSum dataset. Interestingly, ACUEVAL does not show any improvement over BelugaEval on the SUMMEVAL dataset. We emphasize the importance of referring back to the balanced accuracy results in Section 4.3, especially considering the substantial class imbalance present in these datasets.

contain 1800 examples in total.

F Comparison between LLM-based Evaluation Metrics.

The prompt used for BelugaEval can be found in Figure 5. This is very similar to the prompt used for G-Eval except that we change the chain-of-thought prompt to the instruction Fabbri et al. (2021) uses for human annotation. We notice that this more targeted prompt improves the performance. Since we use StableBeluga 2 as the LLM, we use greedy decoding for reliable predictions. We also use the original score normalization technique outlined in Liu et al. (2023a).

G Prompts

G.1 ACUEVAL Prompts

ACU Generation. We show our 5-shot prompt for generating the ACUs in Figure 3. Examples are taken from the ROSE dataset. For 1-shot and 3-shot, we select the first one and first three examples, respectively. For FactScore-style prompt, we use the prompt in Min et al. (2023), which contains multiple human-written in-context examples.

ACU Verification. The prompt is shown in Figure 4. For FactScore-style prompt, we use the provided prompt: "{{ACU}} True or False?"

G.2 Refinement Prompts

For refinement, we use Figure 6 for ACUEVAL-style and Figure 7 for G-Eval-style prompt. We note that the two prompts have the same refinement prompt and differs only in the comment section: ACUEVAL-style comment contains a list of incorrect atomic facts, while the comment with G-Eval-style is a free-form text generated by the scoring model.

H Refinement Examples

We show examples of refinement in Figure 8.

Please breakdown the following passage into independent facts: Has Diamond Blackfan Anaemia so her body can't produce red blood cells . The condition is so rare it affects only around 125 people in the country . Amie has to go into hospital every four weeks for a blood transfusion . For six days out of the week, Amie is attached to a tube which removes excess iron from her blood .
- Amie has Diamond Blackfan Anaemia
- Amie's body can't produce red blood cells
- The condition is so rare
- The condition affects only around 125 people
- only around 125 people in the country is affected
- Amie has to go into hospital
- Amie has to go into hospital every four weeks
- Amie has to go into hospital for a blood transfusion
- Amie is attached to a tube
- For six days out of the week, Amie is attached to a tube
- the tube removes excess iron
- excess iron is removed from Amie's blood.

Please breakdown the following passage into independent facts: Paul Goldstein travels the world photographing animals in their habitats . Chooses his favourite mother-child shots for Mother's Day . Highlights include a newborn giraffe walking within 15 minutes of birth .
- Paul Goldstein travels the world.
- Paul Goldstein photographs animals.
- Animals are in their habitats.
- Paul Goldstein chooses his favourite mother-child shots for Mother's Day.
- Highlights include a giraffe.
- The giraffe is a newborn.
- The giraffe is walking within 15 minutes of birth.

Please breakdown the following passage into independent facts: Paula Deen Cuts The Fat will feature 200 light recipes and low-fat updates to 50 of the chef's 'classic' dishes, she said . Deen also has a new distribution deal with Hachette Book Group . The publisher will release multiple new cookbooks from Deen, and will reissue her previous books in print and electronic form . Just days ago Deen announced she will launch a daily radio show and weekly podcast in May .
- Paula Deen said Cuts The Fat will feature 200 light recipes
- Paula Deen said Cuts The Fat will feature light recipes
- Paula Deen said Cuts The Fat will feature low-fat updates to the chef's 'classic' dishes
- Paula Deen said Cuts The Fat will feature low-fat updates to 50 of the chef's 'classic' dishes
- Deen has a new distribution deal
- Deen has a distribution deal with Hachette Book Group
- Hachette Book Group will release multiple new cookbooks
- Hachette Book Group is a publisher
- Hachette Book Group will release new cookbooks from Deen
- Hachette Book Group will reissue Dean's previous books
- Hachette Book Group will reissue Dean's previous books in print
- Hachette Book Group will reissue Dean's previous books in electronic form
- Just days ago Deen made an announcement
- Deen announced she will launch a daily radio show
- Deen announced she will launch a daily radio show in May
- Deen announced she will launch a weekly podcast
- Deen announced she will launch a weekly podcast in May

Please breakdown the following passage into independent facts: Winston Reid looks set to sign a contract extension at West Ham . West Ham chairman David Gold: 'I am hopeful that Winston will sign for us' Reid is out of contract this summer and had been linked with other clubs . Tottenham Hotspur and Arsenal were among those interested . The New Zealand defender has been at Upton Park since 2010 .
- Winston Reid looks set to sign a contract extension
- Winston Reid looks set to sign a contract extension at West Ham
- West Ham chairman David Gold: 'I am hopeful that Winston will sign for us'
- David Gold: 'I am hopeful that Winston will sign for us'
- David Gold is the West Ham chairman
- Reid is out of contract
- Reid is out of contract this summer
- Reid had been linked with other clubs
- Tottenham Hotspur was among those interested
- Arsenal was among those interested
- Reid has been at Upton Park
- Reid has been at Upton Park since 2010
- Reid is a defender
- Reid comes from New Zealand

Please breakdown the following passage into independent facts: Boy was admitted to hospital in Italy and suffered four seizures . Was suffering from posterior reversible encephalopathy syndrome (PRES) Doctors were baffled as scans ruled out all usual major causes of PRES . His teeth were black so they found he had been eating masses of liquorice . Had consumed double the recommended daily amount of glycyrrhizin . Doctors now calling for safety warning to be printed on sweet packets .
- Boy was admitted to hospital
- Boy was admitted to hospital in Italy
- Boy suffered seizures
- Boy suffered four seizures
- Boy was suffering from posterior reversible encephalopathy syndrome (PRES)
- Doctors were baffled
- Scans ruled out all usual major causes of PRES
- The boy's teeth were black
- Doctor found the boy had been eating liquorice
- Doctor found the boy had been eating masses of liquorice
- The boy had consumed double the recommended daily amount of glycyrrhizin
- The boy had consumed glycyrrhizin
- Doctors now calling for safety warning to be printed
- Doctors now calling for safety warning to be printed on sweet packets

Please breakdown the following passage into independent facts: {{Summary}}

Figure 3: 5-shot prompt for ACU generation.

Read the passage and the statement. Then, answer whether all the information in the statement can be found in the passage.

Passage: {{Document}}

Statement: {{ACU}}

You are ONLY allowed to answer with Yes or No.

Figure 4: Zero-shot prompt for ACU verification.

In this task you will evaluate the quality of summaries written for a news article.

To correctly solve this task, follow these steps:
1. Carefully read the news article, be aware of the information it contains.
2. Read the summary.
3. Rate each summary on a scale from 1 (worst) to 5 (best) by its consistency.

Definition:
The consistency rating measures how well the facts in the summary are consistent with the facts in the original article.
Consider whether the summary does reproduce all facts accurately and does not make up untrue information.

Article:
{{Document}}

Summary:
{{Summary}}

Consistency Score:

Figure 5: Prompt for BelugaEval.

You will be given a document and a summary. You will then be given a comment on the summary.

Your task is to revise the summary given the comment.

Please make sure you address all the suggestions by only making the least amount of changes.

Document:

{{Document}}

Summary:

{{Summary}}

Comment:

The summary is not consistent with the source text. The source text does not mention the following facts:
- {{Incorrect Atomic Fact 1}}
- {{Incorrect Atomic Fact 2}}
...
- {{Incorrect Atomic Fact N}}

The summary should not include information that is not present in the article. Please check the document for the correct information and make appropriate edits.

Revised Summary:

Figure 6: Prompt for correcting faithfulness errors with ACUEval-style comment. {{Incorrect Atomic Fact i}} is replaced with {1...N} incorrect atomic facts that are judged as not consistent with the source.

You will be given a document and a summary. You will then be given a comment on the summary.

Your task is to revise the summary given the comment.

Please make sure you address all the suggestions by only making the least amount of changes.

Document:

{{Document}}

Summary:

{{Summary}}

Comment:

{{Comment}}
The summary should not include information that is not present in the article. Please check the document for the correct information and make appropriate edits.

Revised Summary:

Figure 7: Prompt for correcting faithfulness errors with G-Eval-style comment. The {{Comment}} is replaced with the continuation of G-Eval containing an explanation generated by GPT-4.

Document: It had already been announced the 30-year-old would leave the Warriors when his contract expires this summer. The South Africa-born forward becomes Sale's first new signing for next season. "Josh is a world class ball-carrying forward," said Sharks director of rugby Steve Diamond. "He has been performing well for both Glasgow and Scotland over the past five years and will complement the squad in the ball-carrying department." Strauss joined Glasgow from Super Rugby side Lions in 2012 and, after three years of residency, qualified to play for Scotland in time for the 2015 World Cup. Number eight Strauss has been sidelined by a kidney injury sustained during the Six Nations defeat by France, ruling him out of the rest of the tournament. "Since coming to the UK I always had it in the back of my mind that I wanted to play in the Premiership," said Strauss. "I have really enjoyed my time with the Warriors, but I asked my agent to look around for me. He came back with Sale Sharks who have a good name, a good record in the Premiership and looked an attractive proposition."
Original Summary: Premiership side Sale Sharks will sign Scotland and Glasgow back row forward Josh Strauss on a three-year deal from next season.
Refined Summary w. G-Eval prompt: Premiership side Sale Sharks will sign Scotland and Glasgow back row forward Josh Strauss on a three-year deal from next season.
Refined Summary w. ACUEVAL: Premiership side Sale Sharks will sign Scotland and Glasgow back row forward Josh Strauss from next season.

Document: Real Madrid's La Liga and Champions League chances have been dealt a major blow with confirmation that Luka Modric could miss the rest of the season with a knee injury. The Croatian midfielder pulled up in the second half against Malaga on Saturday and tests on Sunday confirmed a sprained ligament in his right knee with a recovery time of between five and six weeks. The news comes as the club wait to assess the full extent of Gareth Bale's calf injury, picked up in the same game. The connection between the former Tottenham pair is an essential part of Carlo Ancelotti's side's chances of picking up silverware this season. Former Tottenham midfielder Luka Modric is helped off the Bernabeu pitch by medical staff . Cristiano Ronaldo was the first Madrid player to go over to Gareth Bale, who picked up an injury on Saturday . Luka Modric hobbles off the Bernabeu pitch have sprained a ligament in his right knee on Saturday . Modric could still be back for the Champions League final but Real Madrid will have to get there without him. On Wednesday when they play the second leg of their quarter-final against Atletico Madrid, poised at 0-0 after the first 90 minutes, they will almost certainly also be without Bale. It's one year on from his dramatic winner in the Copa del Rey final but Bale, who also scored in the Champions League final and the World Club Cup last year, is likely to be in a race to get back before the Champions League semi-finals on 5-6 May should Real reach them. Gareth Bale went down injured after less than five minutes at the Bernabeu on Saturday evening . Real Madrid will have Karim Benzema back against Atletico so he will partner Crisitano Ronaldo up front with James Rodriguez, Toni Kroos and Isco plus whoever Ancelotti calls upon to replace Modric in a midfield four. German international Sami Khedira and former Real Sociedad player Asier Illaramendi are the options. Bale's absence will be a massive blow for Ancelotti. He said in midweek: 'Bale was fantastic last season and has been very good this year. He scored goals in the finals last season and he will be important for us as this season reaches its climax.' And most Real Madrid supporters see Modric as even more important to the team. The Croatian injured himself towards the end of last year and when he missed three months of football it coincided with the team's worst run of form. The Welsh forward was forced off and faces a race to be fit for Real's Champions League quarter-final . The Spanish press reacts to Real Madrid's injury 'crisis' having beaten Malaga 3-1 on Saturday . Cristiano Ronaldo celebrates after securing the 3-1 victory late on against Malaga on Saturday .
Original Summary: - Luka Modric, the Croatian midfielder, is likely to miss the rest of the season with a knee injury. - Gareth Bale, who also scored in the Champions League final and the World Club Cup last year, is likely to be in a race to get back before the Champions League semi-finals on 5-6 May should Real reach them. - Karim Benzema, the Real Madrid midfielder, is back against Malaga after picking up a injury on Saturday. - James Rodriguez, Toni Kroos and Isco, plus whoever Ancelotti calls upon to replace Modric in a midfield four
Refined Summary w. G-Eval prompt: - Luka Modric, the Croatian midfielder, could miss the rest of the season with a knee injury, but may still be back for the Champions League final. - Gareth Bale, who also scored in the Champions League final and the World Club Cup last year, picked up a calf injury and is likely to be in a race to get back before the Champions League semi-finals on 5-6 May should Real reach them. - Karim Benzema, the Real Madrid midfielder, will be back for the game against Atletico Madrid. - James Rodriguez, Toni Kroos and Isco, plus whoever Ancelotti calls upon,
Refined Summary w. ACUEVAL: - Luka Modric, the Croatian midfielder, is likely to miss the rest of the season with a knee injury. - Gareth Bale, who also scored in the Champions League final and the World Club Cup last year, is likely to be in a race to get back before the Champions League semi-finals on 5-6 May should Real reach them. - Real Madrid will have Karim Benzema back against Atletico, and he will partner with Cristiano Ronaldo up front. - James Rodriguez, Toni Kroos and Isco, plus whoever Ancelotti calls upon, will replace Modric in a midfield four.

Figure 8: Examples of refined summary given G-Eval-style prompt and prompt using ACUEVAL.