Resource-Efficient Reference-Free Evaluation of Audio Captions

Anonymous ACL submission

Abstract

To establish the trustworthiness of systems that automatically generate text captions for audio, images and video, existing reference-free metrics rely on large pretrained models which are impractical to accommodate in resource-constrained settings. To address this, we propose some metrics to elicit the model's confidence in its own generation. To assess how well these metrics replace correctness measures that leverage reference captions, we test their calibration with correctness measures. We discuss why some of these confidence metrics align better with certain correctness measures. Further, we provide insight into why temperature scaling of confidence metrics is effective. Our main contribution is a suite of well-calibrated lightweight confidence metrics for reference-free evaluation of captions in resource-constrained settings.

1 Introduction

Automated context awareness through sensors such as microphones and cameras is being relied on for applications as diverse as home security, military surveillance and machine condition monitoring. When such a system generates unreliable content, the stakes can be high. For example, if a surveillance system mistakenly captions a woodpecker's pecks as gunshots, that could trigger a security threat warning.

The traditional way of judging the quality of generated text is to measure its overlap or similarity with one or more reference texts. This is infeasible when the model is deployed, since reference captions are unavailable. Existing reference-free metrics to evaluate generated text depend on large pretrained models, which occupy too much storage and compute for deployment in resource-constrained settings. Hence, we investigate low-compute methods to evaluate caption quality in the absence of references. Specifically, we contribute the following:

- We propose reference-free evaluation metrics for audio captions that alleviate the need to store and run large pretrained models.
- We validate these metrics by treating them as confidence metrics, and assess their calibration with widely accepted correctness measures.
- We illustrate why temperature scaling of confidences is effective.

2 Related Work

2.1 Evaluating quality of generated text

2.1.1 In the presence of reference text

There are several ways to evaluate the quality of generated text when reference text is available. BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004) measure n-gram overlap, while CIDER (Vedantam et al., 2015) measures the cosine similarities between vectors consisting of TF-IDFs (Jones, 1972) of n-grams. SPICE (Anderson et al., 2016) measures the overlap between scene graphs of reference and generated texts. BERTScore (Zhang et al., 2019), BLEURT (Sellam et al., 2020) and FENSE (Zhou et al., 2022) leverage pretrained language models in an attempt to capture semantic similarities.

2.1.2 In the absence of reference text

In the absence of reference text, evaluating the quality of generated text is more challenging. Often, large pretrained models are used (Fu et al., 2024; Saha et al., 2024; Huang et al., 2024; Jiang et al., 2024; Xu et al., 2023; Qin et al., 2023; Mehri and Shwartz, 2023; Liu et al., 2023; Tian et al., 2023; Zhong et al., 2022; Yuan et al., 2021; Liu et al., 2021; Mehri and Eskenazi, 2020; Pang et al., 2020). The effort to transfer these evaluation capabilities to smaller models (Liu et al., 2024a,b) is nascent. Moreover, these pretrained models have an inherent bias to favor generations from models like them-

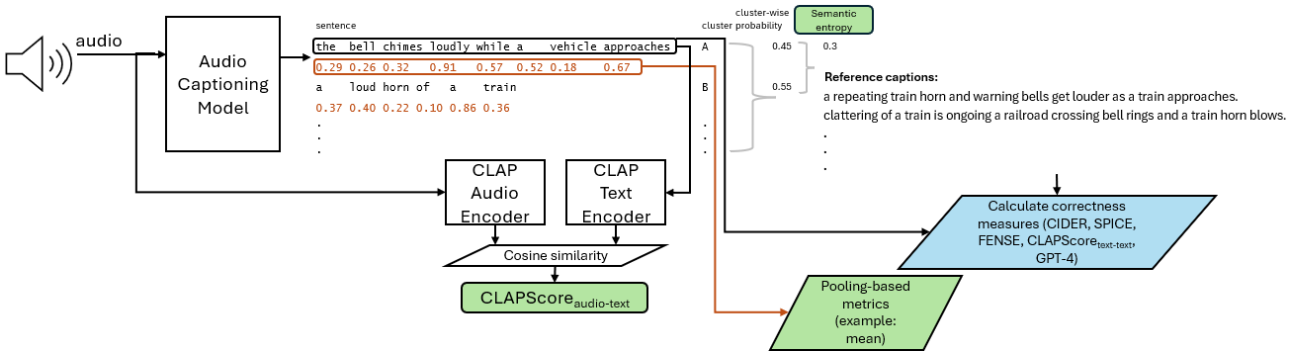


Figure 1: Our framework of obtaining confidence metrics (green) and correctness measures (blue).

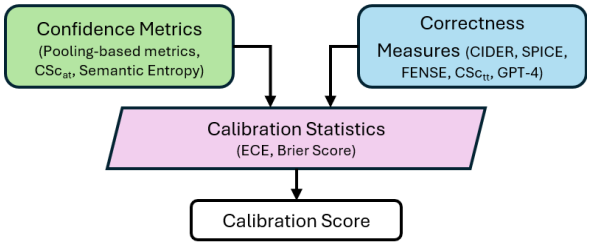


Figure 2: Our framework of measuring calibration of confidence metrics with correctness measures.

selves (Liu et al., 2024c, 2023), and can be biased against higher-quality outputs, including those written by humans (Deutsch et al., 2022). Further, these metrics may rely on spurious correlations with measures such as word overlap, perplexity, and length (Durmus et al., 2022), may be confused by truncation errors, and errors in certain locations in the generation (He et al., 2023). Also, verbalized confidences are not well-calibrated for difficult queries and object counting (Groot and Valdenegro Toro, 2024).

To evaluate the quality of generated text conditioned on other media such as images in the absence of reference text, large pretrained models are again commonly used. To detect hallucinations in image captions,(Petryk et al., 2024) used a Large Language Model (LLM) to extract groundable objects from the captions and measured their semantic similarities with objects detected in the image. Another method is to repeatedly generate an image from the generated caption using a large model followed by captioning the generated image to find a semantic drift indicating lack of coherence (Cao et al., 2024). These large models occupy a huge amount of space and compute, which makes it difficult to deploy them in resource-constrained settings, such as on edge devices. This points to the need to develop a low-compute evaluation metric for captions in the absence of reference text. We explore some options for such a metric, and, by treating them as confidence metrics, assess their alignment with correctness measures that use reference text, in the framework of calibrating confidence metrics with correctness measures.

2.2 Calibration

When a model is subjected to unseen data, a confidence metric is useful to indicate the reliability of the model's output. It is common to measure *calibration* of the confidence metric with correctness measures that rely on the ground truth. One calibration statistic is the Expected Calibration Error (ECE) (Guo et al., 2017), which partitions the n confidences corresponding to n samples into M equally spaced bins $B_1, B_2, ..., B_M$, and then computes a weighted average of the absolute differences between the confidences and the correctnesses in each bin, where the weight is determined by the number of confidences in that bin.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |corr(B_m) - conf(B_m)| \quad (1)$$

$corr(B_m)$ refers to the average correctness of all samples whose confidences belong to B_m, and $conf(B_m)$ refers to the average confidence of all samples whose confidences belong to B_m. Another calibration statistic Brier Score (BS) (Brier, 1950) is the Mean Squared Error between confidence and correctness across all n samples $[x_1, x_2, ..., x_n]$.

$$BS = \frac{1}{n} \sum_{i=1}^{n} (corr(x_i) - conf(x_i))^2 \quad (2)$$

A lower value is better for both calibration statistics.

3 Procedure

We consider an audio captioning model AC which, conditioned on an audio clip a, generates text $t = [t_1, t_2, ..., t_n]$, where t_i is the i^{th} token. Let $p = [p_1, p_2, ..., p_n]$ be the list of respective token probabilities. In this section, we will describe the confidence metrics we developed for audio captions. All of these metrics are deployed during inference and do not require any interference during training. An overall framework diagram is shown in Figure 1.

3.1 Pooling-based metrics

To calculate the confidence of the generated text, we pool probabilities of the generated tokens. We define the arithmetic mean of the token probabilities, henceforth referred to as $AM(t)$ or simply as AM as

$$AM(t) = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{3}$$

We define the geometric mean of the token probabilities, henceforth referred to as $GM(t)$ or simply as GM as

$$GM(t) = (\prod_{i=1}^{n} p_i)^{\frac{1}{n}} \tag{4}$$

The GM can also be viewed as the reciprocal of the perplexity, which is a common measure of the uncertainty of generated text. We refer to the AM and the GM as naive pooling-based confidence metrics. We found that probabilities of non-stopword tokens carry more information, where stopwords refer to frequently occurring words which contribute little semantic value such as a, and, is and the. Hence we also tried pooling using only probabilities of tokens which are among the $noun$, $verb$ and $adjective$ parts of speech, as judged by NLTK (Bird et al., 2009). Formally, let M be the set of all indices identifying tokens from t which are among the $noun$, $verb$ or $adjective$ parts of the speech. We define the selective arithmetic mean of the token probabilities, henceforth referred to as $SAM(t)$ or simply as SAM as

$$SAM(t) = \frac{1}{|M|}\sum_{i \in M} p_i \tag{5}$$

We define the selective geometric mean of the token probabilities, henceforth referred to as $SGM(t)$ or simply as SGM as

$$SGM(t) = (\prod_{i \in M} p_i)^{\frac{1}{|M|}} \tag{6}$$

We refer to the SAM and the SGM as selective pooling-based confidence metrics. The collection of naive pooling-based confidence metrics and selective pooling-based confidence metrics is referred to as pooling-based metrics.

3.1.1 Temperature Scaling

We optionally apply temperature scaling to the list of logits $q = [q_1, q_2, ..., q_n]$ corresponding to token probabilities before the softmax layer. To clarify, the relationship between p and q is $p = softmax(q)$. Using a scalar temperature $temp$, the temperature-scaled probabilities $p'(temp) = [p'_1(temp), p'_2(temp), ..., p'_n(temp)]$ are obtained from q as follows:

$$p'_i(temp) = \frac{e^{\frac{q_i}{temp}}}{\sum_{q_j \in q} e^{\frac{q_j}{temp}}} \tag{7}$$

3.2 CLAPScore

To measure how similar the generated text t is to the audio a, we measure the cosine similarity between them in the multimodal space enabled by the CLAP (Elizalde et al., 2023) training mechanism. The $CLAPScore_{at}$ or CSc_{at} is defined as

$$CSc_{at}(a,t) = \frac{ad\_emb(a).tx\_emb(t)}{||ad\_emb(a)||||tx\_emb(t)||}, \tag{8}$$

where ad_emb and tx_emb are both unary functions that project their audio and text inputs respectively into a shared multimodal space.

3.3 Semantic Entropy

To measure how consistent the model's responses are across generations for the same input, we adapted the concept of semantic entropy (Farquhar et al., 2024) for audio captions. For an audio clip a, we sample a set of p generations $T = t^{(1)}, t^{(2)}, ..., t^{(p)}$. For two text generations q and r, let us define the $CLAPScore_{tt}$, abbreviated as CSc_{tt} as

$$CSc_{tt}(q,r) = \frac{tx\_emb(q).tx\_emb(r)}{||tx\_emb(q)||||tx\_emb(r)||} \tag{9}$$

Using CSc_{tt} as the distance metric, we perform agglomerative clustering within T which stops when the minimum distance between clusters exceeds a certain threshold $h \in [-1, 1]$. In our experiments, $h = 0.7$. Next, for the l^{th} cluster c_l, we calculate its probability $P(c_l)$ as the average of probabilities of all its generations, where the probability

of a generation is simply the average of all token probabilities.

$$P(c_l) = \frac{1}{|c_l|} \sum_{t^{(j)} \in c_l} \frac{1}{|t^{(j)}|} \sum_{i=1}^{|t^{(j)}|} t_i^{(j)}, \quad (10)$$

where $|c_l|$ is the number of generations in c_l, and $|t^{(j)}|$ is the number of tokens in $t^{(j)}$. To obtain a valid probability distribution P' that sums to one, we normalize P using the L^1-norm. For every $c_i \in T$,

$$P'(c_i) = \frac{P(c_i)}{\sum_{c_j \in T} |P(c_j)|}. \quad (11)$$

Finally, we calculate the semantic entropy $SE(T)$, also referred to as SE as

$$SE(T) = -\sum_{c_i \in T} P'(c_i) log(P'(c_i)) \quad (12)$$

In our experiments, $p = 7$. To stay consistent with the property among confidence metrics of being constrained between 0 and 1, with higher being better, we define the Inverse Semantic Entropy $ISE(T)$, also referred to as ISE as

$$ISE(T) = 1 - min(SE(T), 1). \quad (13)$$

The inverse refers to an additive inverse. It seemed reasonable to clip the SE to 1 because empirically, the SE being higher than 1 is rare (only happens for less than 2% captions in the validation set), and the semantic instability of such captions is well-highlighted by the ISE even after clipping.

4 Experiments

In this section, we describe the experiments to measure how well the confidence metrics we described in Section 3 calibrate with correctness measures, also shown in Figure 2.

Correctness Measures: Apart from the traditional correctness measures CIDER and SPICE and the pretrained model-based correctness measure FENSE, all of which were introduced in Subsubsection 2.1.1, we use two other correctness measures to judge the correctness of the generated text with respect to a reference text. The first new correctness measure is the the $CLAPScore_{tt}$ or CSc_{tt} defined in Equation 9, which calculates the cosine similarity between the generated text and the reference text in the audio-text multimodal space. The second is GPT-4's judgment regarding how well the generated text describes the audio which is described by the reference text. To study the relationships of correctness measures with each other, we calculated the Pearson correlations between them.

Model architecture: Following (Mei et al., 2021), our audio captioning model consists of a CNN10 PANN encoder (Kong et al., 2020) followed by four layers of a transformer decoder with two heads each, with a hidden size of 256 and a feedforward dimension of 2048. It uses text embeddings from the bert-L12-H256 model(Turc et al., 2019). This model has about 15 million parameters, and took about 120 GPU hours to train. To test the applicability of our results to other models, we also used an alternate model with a hidden size of 128 instead of 256, which uses text embeddings from the bert-L12-H128 model(Turc et al., 2019). This model had about 8 million parameters.

Datasets: We trained this model with our own audio captioning dataset (details in Appendix A). For evaluation, we used the evaluation splits of the AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) datasets, which have 957 and 1045 samples respectively. To find the optimum temperature for calibration, we used the validation splits of these datasets, which have 495 and 1045 samples respectively.

Measuring Calibration: The ECE and the Brier Score are used to measure the calibration of our confidence metrics with correctness measures. For pooling-based metrics, we also test the effectiveness of temperature scaling in improving calibration by selecting from $temp \in 0.1, 0.2, ..., 2.0$ using the validation split.

5 Results

5.1 Identifying clusters in correctness measures

Pearson correlations among correctness measures are shown in Figure 4 for the evaluation split of the AudioCaps dataset. We observe that the traditional correctness measures CIDER and SPICE correlate with each other, while the model-based correctness measures FENSE, CLAPScore_{tt} and GPT-4 correlate with each other. The same trend is observed for the Clotho dataset.

5.2 Evaluating Confidence Metrics

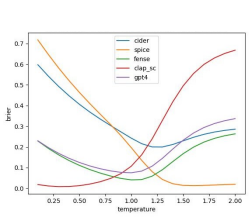
Table 1 shows calibration scores for both datasets using the Brier Score and ECE, when no temperature scaling is used. We observe that pooling-based confidence metrics align well with all correctness

	Brier Score (↓)					Expected Calibration Error (↓)				
AudioCaps										
	CIDER	**SPICE**	**FENSE**	**CSc$_{tt}$**	**GPT-4**	**CIDER**	**SPICE**	**FENSE**	**CSc$_{tt}$**	**GPT-4**
AM	0.24	0.2	**0.04**	0.11	**0.08**	0.21	0.42	0.09	0.31	**0.08**
SAM	0.22	0.15	0.05	0.16	0.09	0.16	0.36	0.11	0.37	0.11
GM	0.22	0.16	**0.04**	0.14	0.08	0.16	0.37	**0.08**	0.36	0.09
SGM	**0.2**	**0.12**	0.05	0.19	0.1	**0.11**	**0.32**	0.12	0.41	0.14
CSc$_{at}$	0.3	0.31	0.06	**0.04**	**0.08**	0.34	0.55	0.15	**0.18**	0.12
ISE	0.6	0.71	0.25	0.05	0.25	0.62	0.82	0.44	**0.18**	0.41
Clotho										
AM	0.2	0.21	0.05	0.1	0.08	0.35	0.45	0.12	0.29	0.11
SAM	0.15	0.16	0.05	0.15	0.08	0.28	0.38	0.11	0.35	0.1
GM	0.16	0.16	**0.04**	0.14	**0.07**	0.28	0.38	**0.09**	0.35	**0.07**
SGM	**0.12**	**0.12**	0.05	0.19	0.08	**0.22**	**0.32**	0.11	0.41	0.11
CSc$_{at}$	0.31	0.36	0.1	**0.03**	0.12	0.49	0.59	0.25	**0.14**	0.23
ISE	0.65	0.74	0.32	0.06	0.33	0.73	0.83	0.5	0.21	0.49

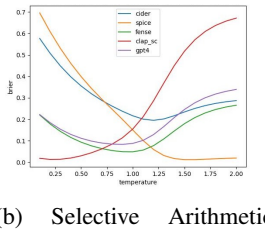
Table 1: Calibration scores on the evaluation splits of AudioCaps and Clotho with no temperature scaling.

	CIDER	**SPICE**	**FENSE**	**CSc$_{tt}$**	**GPT-4**	**Avg w/o TS**	**Avg w/ TS**
AudioCaps							
AM	.2	.014	**.041**	**.008**	**.075**	.132	.068
SAM	.202	**.012**	.048	.013	.083	.131	.072
GM	.193	.016	**.041**	.009	**.075**	.127	.067
SGM	**.191**	.013	.049	.014	.084	.131	.07
Clotho							
AM	.08	**.007**	.044	**.01**	**.069**	.128	.042
SAM	.079	**.007**	.049	.017	.077	.118	.046
GM	.076	**.007**	**.042**	.012	**.069**	.113	.041
SGM	**.073**	**.007**	.05	.02	.076	.11	.045

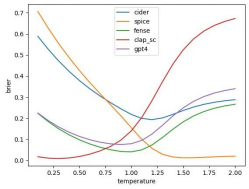
Table 2: Brier scores (↓) on the evaluation splits of AudioCaps and Clotho when using Temperature Scaling (TS).



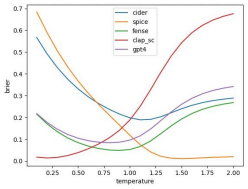
(a) Arithmetic Mean



(b) Selective Arithmetic Mean



(c) Geometric Mean



(d) Selective Geometric Mean

Figure 3: Brier scores over temperatures for the AudioCaps dataset. Each plot shows the variation of all correctness measures over temperatures for a single confidence metric.

	CIDER	**SPICE**	**FENSE**	**CLAPScore$_{tt}$**	**GPT-4**
CIDER	1	0.56	0	0.02	0
SPICE	0.56	1	-0.06	-0.07	-0.07
FENSE	0	-0.06	1	0.69	0.64
CLAPScore$_{tt}$	0.02	-0.07	0.69	1	0.77
GPT-4	0	-0.07	0.64	0.77	1

Figure 4: Pearson correlation between correctness measures for the AudioCaps dataset.

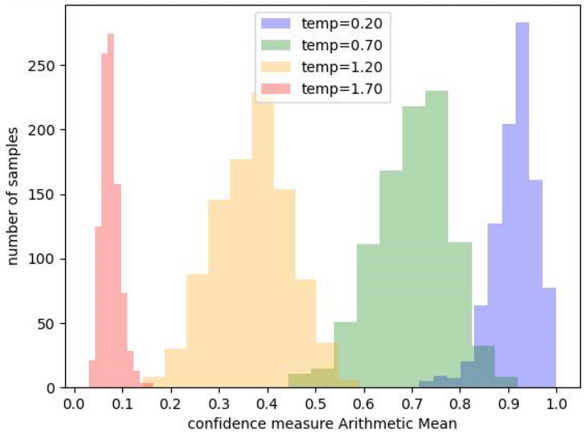


Figure 5: Variation of distribution over temperatures of the Arithmetic Mean confidence metric for the AudioCaps dataset.

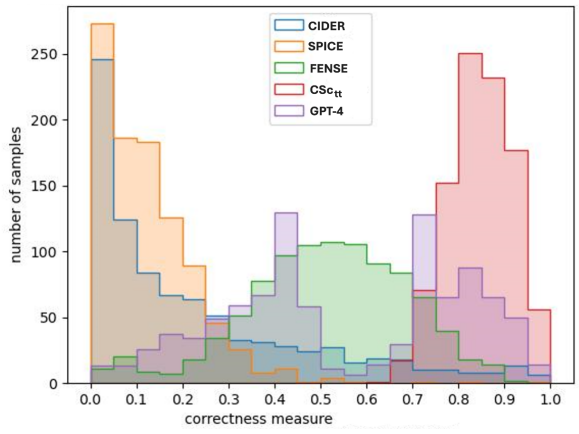


Figure 6: Distributions of correctness measures for the AudioCaps dataset.

measures. For traditional correctness measures CIDER and SPICE, selective pooling achieves a clear improvement over conventional pooling. This may be because CIDER gives less weight to more frequently occurring tokens, which are also the ones ignored by selective pooling, and the SPICE metric also ignores stopwords, except when needed to determine relations. For model-based correctness measures FENSE, CLAPScore$_{tt}$ and GPT-4, pooling-based confidences continue to perform well. However, using selective pooling does not have an advantage here because model-based metrics look at the entire sentence, part of which we are discarding when using selective pooling.

For the CSc$_{tt}$ correctness measure, we can achieve even better calibration using the CSc$_{at}$ and ISE confidence metrics. This may be because just like the CSc$_{tt}$ correctness measure is lenient in allowing acoustically similar but semantically distant cross-triggers (examples: machine whirring and helicopter flying, typing and object clattering), the CSc$_{at}$ also forgives such cross-triggers because the cosine similarity is measured in the audio-text multimodal space. The ISE also overlooks such cross-triggers because the CSc$_{tt}$ is used to judge consistency between the model's responses, while forming clusters to calculate the entropy.

To demonstrate the generalizability of these results to other models, Table 5 in Appendix B shows the same results when using the alternate model.

5.3 Effect of Temperature Scaling

The curves of calibration quality over temperatures are shown in Figure 3 for the validation split of the AudioCaps dataset. The optimal temperature to calibrate a particular confidence metric with a correctness measure stays the same for both datasets, indicating its generalizability to unseen data.

Table 2 shows calibration results at these optimal temperatures, on the evaluation splits of both datasets. The effectiveness of applying temperature scaling is quite pronounced, as evident from the last two columns of the table which show the calibration scores averaged over correctness measures before and after temperature scaling respectively. The average Brier score for each confidence metric almost halves after using temperature scaling. It is however important to remember that for such a dramatic improvement in calibration to be achieved, a validation set is needed to carefully select the optimal temperature. In cases where such a validation set is not available, selective pooling-based confi-

dence metrics are still the best choice to calibrate well with traditional correctness measures. No one pooling-based metric stands out in performance if temperature scaling is applied, an explanation for which is provided in the next Subsubsection.

5.3.1 Why does temperature scaling work?

Figure 6 shows the distributions of correctness measures, while Figure 5 shows how the distribution of a representative pooling-based confidence metric AM changes over temperature. A low temperature causes the AM to shift to the right, which matches most closely with CSc_{tt}, explaining why a low temperature is needed to calibrate well with CSc_{tt}. Similarly, a high temperature causes the AM to shift to the left, resulting in a distribution similar to those of CIDER and SPICE, explaining why a high temperature is needed to calibrate well with theses two correctness measures. Finally, a moderate temperature allows the confidence metrics's distribution to be centered around 0.5, which matches the distributions of FENSE and GPT-4, explaining why a temperature close to 1 is reasonable for calibrating with these correctness measures.

This ability of pooling-based metrics to adjust their distributions to match with those of correctness measures somewhat compensates for the differences in their computations, resulting in all of them being comparably effective.

6 Conclusion

We propose some resource-efficient reference-free evaluation metrics for audio captions, and validate their effectiveness by measuring their calibration with correctness measures that use references. Finally, we demonstrate the effectiveness of temperature scaling and explain why it is effective. Our work enables the reliable deployment of audio captioning systems in resource-constrained settings.

7 Discussion

Feasibility of deploying our metrics in a resource-constrained setting: The CSc_{at} confidence metric uses GPT-2 (126.38M) as the text encoder and HTSAT (159.45M) as the audio encoder, which results in the need to store and forward-propagate through 159.45M parameters. The ISE needs only the GPT-4 text encoder. This is much lesser than sizes of models from comparable past work that have billions of parameters (Saha et al., 2024; Huang et al., 2024; Liu et al., 2023; Tian et al., 2023; Pang et al., 2020; Xu et al., 2023; Jiang et al., 2024).

Generalizability of our methods to other modalities: The applicability of our proposed reference-free evaluation metrics may extend to captioning systems of other modalities as well. The idea of using as a confidence metric the cosine similarity between embeddings of the conditioning and generated modalities may be extended to the image captioning and video captioning areas. The cosine similarities between CLIP (Radford et al., 2021) embeddings of pairs of images and generated text may be a good indicator of the confidence of the generated image captions. The ISE calculated by using cosine similarities between CLIP text embeddings as the clustering criterion may also be a valuable confidence metric for image captioning models.

Cross-triggers: Expecting an audio captioning model to distinguish between acoustically similar sounds may be unfair, in which case, the CSc_{tt} is the appropriate correctness measure, and the CSc_{at} and the ISE are the recommended confidence metrics. However, if end users of captioning models are unwilling to tolerate cross-triggers, pooling-based confidence metrics are more suitable. To reduce cross-triggers, integrating information from other sensors like cameras and motion sensors can help enhance the system's awareness. However, due to the trade-off between using more sensors and preserving privacy, there is still value in systems that use less sensors.

8 Limitations

- Given that our reference-free evaluation metrics were validated with respect to the existing evaluation metrics that leverage references, our validation is limited by the quality of the existing evaluation metrics and by the quality of the human-written captions that these evaluation metrics depend on. Studying the alignment of these proposed reference-free evaluation metrics with human preferences is beyond the scope of this work.

- The Expected Calibration Error and Brier Score are well-suited to measure the quality of calibration of confidences for classification tasks. Its suitability to measure calibration of natural language is yet to be evaluated independently.

- The potential risk of this work is that our proposed reference-free evaluation metrics, if not providing a true measurement of the confidence of the caption because of the two limiting factors mentioned above, may provide a false sense of reliability.
- Since the objective of the study was not to evaluate the quality of the captioning model, we performed experiments with only a subset of all possible models for the audio captioning task. It is possible, though unlikely, that these results may be less applicable to other model architectures for the same task.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Lele Cao, Valentin Buchner, Zineb Senane, and Fangkai Yang. 2024. Introducing GenCeption for multimodal LLM benchmarking: You may bypass annotations. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. a. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, and Ramakrishna Vedantam. b. coco-caption.

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*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740.

Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. 2022. Spurious correlations in reference-free evaluation of text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. msclap.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

Tobias Groot and Matias Valdenegro Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

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

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13441–13460, Miami, Florida, USA. Association for Computational Linguistics.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. TIGERScore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating Captions for Audios in The Wild. In *NAACL-HLT*.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Etienne Labbé. 2024. aac-metrics.

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.

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2024a. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

Weize Liu, Guocong Li, Kai Zhang, Bang Du, Qiyuan Chen, Xuming Hu, Hongxia Xu, Jintai Chen, and Jian Wu. 2024b. Mind's mirror: Distilling self-evaluation capability and comprehensive thinking from large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

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

Ye Liu, Wolfgang Maier, Wolfgang Minker, and Stefan Ultes. 2021. Naturalness evaluation of natural language generation in task-oriented dialogues using BERT. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, Held Online. INCOMA Ltd.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024c. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics.

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

Shuhaib Mehri and Vered Shwartz. 2023. Automatic evaluation of generative models with instruction tuning. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*.

Xinhao Mei, Qiushi Huang, Xubo Liu, Gengyun Chen, Jingqian Wu, Yusong Wu, Jinzheng Zhao, Shengchen Li, Tom Ko, H Lilian Tang, et al. 2021. An encoder-decoder based audio captioning system with transfer and reinforcement learning for dcase challenge 2021 task 6. *DCASE2021 Challenge, Tech. Rep, Tech. Rep*.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

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

Suzanne Petryk, David Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph Gonzalez, and Trevor Darrell. 2024. ALOHa: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2023. T5Score: Discriminative fine-tuning of generative evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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

Zelin Zhou, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kenny Q Zhu. 2022. Can audio captions be evaluated with image caption metrics? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 981–985. IEEE.

A Experimental Details

A.1 Dataset

Our audio captioning dataset was collected using the same crowdsourcing method as (Kim et al., 2019) by asking people to listen to an audio clip and to write one full English sentence describing its contents. Annotators were instructed to not include names or any personally identifiable information, and were also instructed to avoid offensive language. The dataset has 80,000 audio clips of length 10 seconds, and three captions corresponding to each clip, which were written by three different people. Some examples of captions from our dataset are shown in Table 3.

A.2 Confidence and Correctness Measures

To calculate parts of speech for selective pooling metrics, the $tag.pos\_tag$ function from NLTK version 3.8.1 (Apache License, Version 2.0) was used. The CIDER and SPICE implementations from the pycocoevalcap library (Chen et al., b) were used. To calculate FENSE, we used the 'paraphrase-TinyBERT-L6-v2' model (Reimers and Gurevych, 2019) which is default in the aac-metrics toolkit (Labbé, 2024) (MIT License). To calculate the CLAPScore$_{at}$ and CLAPScore$_{tt}$, we used the '2023' configuration of the CLAP model from the msclap library (Elizalde et al.) (MIT License), which uses GPT-2 (Radford et al., 2019) as the text encoder and HTS-AT (Chen et al., a) as the audio encoder. The prompt to GPT-4 for judging the correctness of a caption with respect to a reference is shown in Table 4. The example scores were calculated using cosine similarities between the 'all-MiniLM-L6-v2' SentenceBERT embeddings.

B Results with Alternate Model

Table 5 shows calibration scores for the evaluation splits of both datasets using the Brier Score and ECE, when no temperature scaling is used, when the alternate model is used.

Captions
A series of beeps from multiple different alarms. A continuous sharp blares of a siren followed by a loud honks and horns. A vehicle with a siren is honking.
Some rustling and a person's grunting and shouting. Someone is coughing loudly and a person suddenly shouts. A woman blows sneezes and shouts.
Metals are continuously screeching. Screeching of an operating machine. Buzzing of an electric device.
A dog howls and barks as a wind instrument is playing. Dog weeping and barking while instrumental music is playing. Musical instrument playing and a dog barking and wailing.
A loud rumble of thunder as the rain falls down. Thunder and heavy rain. A heavy rainfall accompanied by a loud bang of the thunder.
A sound of an mechanical equipment tools. A machine buzzing deeply. Screeching of an operating machine.
A loud screaming shouting and cheering of people. People are shouting and clapping. The people are cheering at full blast.
A baby crying and continuous buzzing of an electronic device. A baby crying constantly and some crackling. A baby is incessantly crying.
A man snores loudly as water rushes. The water is running and the person is snoring. A person snores loudly and water starts to flow.
A loud honking of a train that is passing by. The honking horn of a series of railroad cars moving as a unit by a locomotive or by integral motors. Many cars are making loud horn noises.
Birds are tweeting and chirping simultaneously. Birds singing and whistling wonderfully. A bird is chirping and a whistle can be heard while an equipment is creating a humming sound.
Chime of a musical instrument. The bells are ringing simultaneously. A series of loud chimes and clanks of bells.

Table 3: Example captions from our audio captioning dataset.

You will be given five reference sentences to describe an audio scene, and a new sentence. Using that, please evaluate how well a new sentence describes the audio scene, and provide a score between 0 and 1. Please provide only the score, and no other text. Here are some examples:
Example 1: Reference sentences: people are singing and laughing a person is singing in melodic music while surrounded by a passing vehicle a person is singing while a man is laughing a splashing of water the wind is blowing and vehicles are passing by people are singing while cars pass by and a man in laughing people are laughing and singing while vehicles are passing by New sentence: a person is singing while the children are playing Score: 0.548 **Example 2:** Reference sentences: music is playing a musical effect is playing there is instrumental music playing someone is playing a musical instrument instrumental music is playing New sentence: a musical instrument is playing Score: 0.801 . . . Now it's your turn.

Table 4: Prompt provided to GPT-4 to judge the correctness of a caption with respect to a reference.

	Brier Score (↓)					**Expected Calibration Error (↓)**				
AudioCaps										
	CIDER	**SPICE**	**FENSE**	**CSc**$_{tt}$	**GPT-4**	**CIDER**	**SPICE**	**FENSE**	**CSc**$_{tt}$	**GPT-4**
AM	0.20	0.12	**0.06**	0.16	**0.08**	0.18	0.32	**0.06**	0.39	0.12
SAM	0.20	0.10	**0.06**	0.19	0.10	0.13	0.29	0.09	0.42	0.15
GM	0.19	0.09	**0.06**	0.20	0.10	0.14	0.28	**0.06**	0.43	0.16
SGM	**0.18**	**0.08**	0.07	0.23	0.11	**0.09**	**0.25**	0.10	0.46	0.19
CSc$_{at}$	0.32	0.30	0.11	**0.04**	**0.08**	0.37	0.54	0.23	**0.18**	**0.11**
ISE	0.46	0.46	0.27	0.16	0.23	0.44	0.58	0.38	0.24	0.30
Clotho										
AM	0.13	0.14	0.05	0.15	**0.07**	0.26	0.35	0.08	0.37	**0.06**
SAM	0.12	0.12	0.05	0.18	**0.07**	0.23	0.32	0.09	0.40	0.10
GM	0.10	0.10	**0.04**	0.19	**0.07**	0.21	0.30	**0.07**	0.42	0.07
SGM	**0.09**	**0.09**	0.05	0.23	0.08	**0.17**	**0.26**	0.10	0.46	0.12
CSc$_{at}$	0.30	0.35	0.12	**0.03**	0.12	0.49	0.58	0.27	**0.14**	0.23
ISE	0.40	0.48	0.25	0.17	0.25	0.50	0.59	0.37	0.25	0.34

Table 5: Calibration scores on the evaluation splits of AudioCaps and Clotho with no temperature scaling with the alternate model.