Large Language Models as a Proxy For Human Evaluation in Assessing the Comprehensibility of Disordered Speech Transcription

Katrin Tomanek [1] Jimmy Tobin [1] Subhashini Venugopalan [1] Richard Cave [1 2] Katie Seaver [1] Rus Heywood [1] Jordan Green [1 3]

Abstract

Automatic Speech Recognition (ASR) systems, despite significant advances in recent years, still have much room for improvement particularly in the recognition of disordered speech. Even so, erroneous transcripts from ASR models can help people with disordered speech be better understood. Evaluating the efficacy of ASR for this use case requires a methodology for measuring the impact of transcription errors on the intended meaning and comprehensibility. Human evaluation is the gold standard for this, but it can be laborious, slow, and expensive. Here, we tuned and evaluated large language models (LLMs) and found them to be a better proxy for human evaluators compared to typical sentence similarity metrics. We further present a case-study of using our approach to make ASR model deployment decisions in a live video conversation setting.

1. Introduction

Automatic Speech Recognition (ASR) systems have great potential to improve communication for people with speech impairments. Although recent advances in personalized ASR have significantly improved the performance of these systems (Green et al., 2021), the impact of transcription errors on functional communication is highly variable. Even when Word Error Rates (WER) are relatively low, the semantic changes introduced by the errors can critically affect the intended meaning. Although the efficacy of ASR systems for improving functional communication depends on their ability to preserve meaning, these systems are typically evaluated primarily based on WERs (Cave & Bloch, 2021) and sometimes using semantic similarity measures e.g. (Tobin et al., 2022; Kim et al., 2021). WERs, however, do not accurately convey the level of comprehensibility because listeners can reconstruct a message using various contextual cues gleaned from their knowledge of semantics, syntax, pragmatics, and (when available) nonverbal communication. This is more pronounced when measuring comprehensibility in persons with impaired speech (Pommée et al., 2022).

The gold standard for measuring comprehensibility of transcripts for such use cases is judgments made by human evaluators (Hustad, 2008; Yorkston et al., 1996). However it is also a costly and time intensive process, and vulnerable to a variety of rater biases such as the listener's age (Dagenais et al., 2011) or familiarity with the speaker or the spoken content (D'Innocenzo et al., 2006). Here, we look at a fast and less expensive automated proxy measurement for comprehensibility in order to improve deployment decisions, especially around personalized ASR models for speakers with impaired speech.

Related Work. Previous works have investigated using metrics from the NLP literature e.g. ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) and even LM-based semantic similarity measures like BERTScore (Zhang* et al., 2020) or SemDist (Kim et al., 2021) to assess ASR model performance. Similarly there have also been metrics based on word embeddings, but those have also been found to be unreliable (Zhou et al., 2022) and in particular lack nuance of meaning. Overall, the LM-based metrics were found to be more robust to errors like normalization and contraction errors than simply using BLEU scores or WER (Tobin et al., 2022).

The underlying hypothesis of this work is that recent Large Language Models (LLMs), which have achieved SOTA performance on a large variety of NLP tasks (gpt), can be more effective at measuring text-based comprehensibility. Inspired by recent works that have effectively prompt-tuned LLMs (Lester et al., 2021) on small amounts of data (Mozes et al., 2023), we prompt-tune and evaluate LLMs on their ability to classify and predict human judgements on whether a transcription's meaning is preserved. We then introduce a metric (**LATTEScore**) that estimates the utility of an ASR

[1] Google Research, USA [2] Language and Cognition, UCL, UK [3] MGH Institute of Health Professions, USA. Correspondence to: Katrin Tomanek <katrintomanek@google.com>, Jordan Green <jgreen2@mghihp.edu>.

AI & HCI workshop at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

model as the percentage of transcriptions for which the meaning is preserved. In a case study on conversational speech, we find our score to be most correlated with human judgements in selecting personalized ASR models.

2. Transcript Comprehensibility Data Set

Severity	Mild	Moderate	Severe
# Speakers	13	15	14

Table 1. Distribution of speakers across impairment severities.

We created a dataset of tuples of ground truth and ASR transcript along with human assessment of the level of meaning change induced by transcription errors. To do so, we follow a procedure similar (Tobin et al., 2022) and obtain assessments from Speech Language Pathologists (SLPs) on erroneous[1] transcriptions on a 3 point scale (Table 3). A rating of 0 was used to indicate meaning was fully preserved, 1 to indicate there are errors but most of the meaning is still preserved, and 2 to indicate meaning was significantly lost. In accordance with ethical data collection practices, the utterances in this dataset were contributed voluntarily by paid participants who consented to recording extemporaneous speech for research purposes[2].

Although the 3 point scale is easier for SLPs to capture nuances when grading the semantic error severity, this can be mapped to a binary objective: Transcripts with error severity at levels 0 and 1 are considered to be "recoverable," i.e., a communication partner would be able to understand what a speaker meant to say. An error level 2 represents a communication breakdown, where the perceived meaning is lost or drastically altered from the speaker's intent.

Based on applying an ASR model[3] on speech samples from 42 speakers with different levels of speech impairments, we collected an English dataset of $4,731$ prompted and unprompted ground truth and erroneous transcription pairs and obtained SLP annotations on these pairs as described above (see Tables 3 and 1). Overall, in 43% of the cases, the meaning was preserved, while 57% of the transcriptions exhibited a true meaning loss compared to the speaker's intent. On a subset of $\approx 5\%$ of the examples, we obtained annotations from 2 different SLPs and used this to measure the Inter-Annotator Agreement (Cohen, 1960). On the 3-way assessment we obtained a $\kappa = 0.64$, however on the merged, binary "meaning preserved" assessment the κ value was higher at 0.70.

Split	train	test	dev
# Examples	2840	940	951
Meaning Preserved	43.6%	40%	45.2%

Table 2. Data splits and percentage of transcripts categorized as *meaning preserved.*

3. Predicting Transcript Comprehensibility

Motivated by the use case of selecting models that help people with impaired speech to be better understood, our goal was to predict for a pair of ground truth and ASR transcription, whether the transcription preserved the meaning of the ground truth or not.

Data split. We split our data set into train (60%), test (20%) and dev (20%) portions such that there is no overlap in ground truth sentence between the splits. Table 2 shows the respective sizes as well as the percentage of examples where meaning was preserved.

Approach overview and metrics. We tested several techniques with different levels of semantic representation for this problem: (1) A logistic regression model based on WER and BERTScore, (2) A logistic regression model with sentence embedding based cosine similarity as feature, and (3) prompt-tuned LLM-based classifiers. For all approaches, we use the dev split to chose best hyper parameters and to select best checkpoints. All results are reported on the test sets as AUC-ROC scores.

Word Error Rate and BERTScore WER is the standard metric used to compare performance of ASR models, however it does not capture semantic similarity. Previous works (Tobin et al., 2022; Kim et al., 2021) have found alternate text generation metrics like BERTScore (Zhang* et al., 2020) to work better when determining error severity. Inspired by this, our baseline was a logistic regression (LR) model that combined BERTScore and WER (split into Deletion Rate, Insertion Rate, Substitution Rate) as features.

SentT5 sentence embedding similarity Cosine similarity of sentence embeddings is another common strategy used to measure semantic similarity of sentences (STS). We use the 11B parameter SentenceT5 model (Ni et al., 2021) to obtain embeddings for the ground truth sentence and ASR transcriptions. We then train a logistic regression classifier that takes as input the cosine similarity between the embeddings, and the length of the two sentences, and predicts whether meaning is preserved.

Large Language Models as classifiers. We used two recent large language models (LLMs) which have been instruction-finetuned on a large number of tasks and ex-

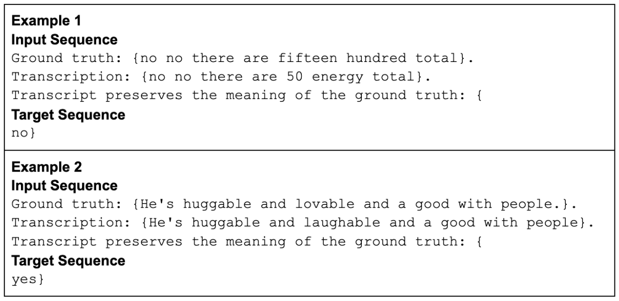
[1] Note that we define "erroneous" transcripts as such where the WER between ground truth and transcript is > 0 (after simple normalization including lower casing and punctuation removal.

[2] A data quarantine process was used to remove utterances containing personally identifiable information (PII) prior to being released for research.

[3] Several different ASR models were used to transcribe speech, the majority being personalized models designed to run on device.

Error Severity	Meaning Preserved	Description	# Examples (%)	Example
0	yes	Meaning is completely preserved	900 (19%)	**G:** I would be fascinated to know your answers. **T:** I *will* be fascinated to know your answers.
1	yes	Some errors, but meaning is mostly preserved.	1145 (24%)	**G:** Yeah I have one basically every day. **T:** Yeah I have *I'm* basically every day.
2	no	Major errors, significant loss of intended meaning.	2686 (57%)	**G:** How large is that file? **T:** How large is a *funnel*?

Table 3. Error severity assessment response scale, descriptions, counts and proportion of the total 4,731 erroneous transcripts with representative examples (**G** for ground truth and **T** for transcript).



Example 1
Input Sequence

```
Ground truth: {no no there are fifteen hundred total}.
Transcription: {no no there are 50 energy total}.
Transcript preserves the meaning of the ground truth: {
```

Target Sequence

```
no}
```

Example 2
Input Sequence

```
Ground truth: {He's huggable and lovable and a good with people.}.
Transcription: {He's huggable and laughable and a good with people}.
Transcript preserves the meaning of the ground truth: {
```

Target Sequence

```
yes}
```

Figure 1. Representation of task for LLM-based classifiers.

hibit strong performance on a variety of benchmarks: (1) *FLAN-T5 XXL* – the 11B parameter encoder-decoder model from (Chung et al., 2022) and (2) *LLM62B* – an in-house 62B parameter decoder-only model.

Our initial few-shot approach to predict scores led to highly variable performance which we exclude from discussion as this has been noted in several prior works (Zhao et al., 2021). We then tune these models via prompt-tuning (Lester et al., 2021), a parameter-efficient tuning (PET) approach where only a small amount of tuneable parameters is prepended to the embedding layer. Standard gradient descent is applied, but only the tuneable parameters are updated while the rest of the model remains frozen. PET approaches are preferable as they are more efficient when it comes to storing and serving checkpoints. In addition, it has recently been shown (Mozes et al., 2023) that these methods also achieve good performance on small training data sets.

Prompt-tuning details. We provide both the ground truth and the ASR transcript as input sequences to the LLM. We prompt-tune the model to predict the labels "yes" or "no" to indicate whether the meaning has been preserved or not. Figure 1 shows how the task is presented to the LLM. We initialize the soft prompts (tuneable parameters) with a random sample of vocabulary token embeddings from the respective model's 5,000 most frequent tokens (similar to the procedure described in (Lester et al., 2021)). We use a prompt length of 5 tokens. We train with a warm-up learning rate schedule with 500 warm up steps to a peak of 0.1 followed by linear decay. We use small batch sizes of 16 for training and limit training to $10k$ steps. We chose the best checkpoint based on the accuracy on the dev set.

For inference, we generate classification scores $s \in [0, 1]$ by obtaining the LLM's log perplexity scores for the tokens corresponding to the two class labels ("yes" and "no"), apply softmax, and then take the score of class "yes".

4. Results

Table 4 shows AUC-ROC scores for the four different approaches, both calculated on the overall test set, and also for different slices of the test set (split by speech type and impairment severity). The LLM-based classifiers consistently outperform other approaches across all slices, with the larger model (62B) clearly surpassing the smaller (11B) one.

The results show a clear ranking of performance as we increase the size and consequently the level of semantics included in the classifier. The LR classifier using WER and BERTScore has the least semantic information and scores lowest. Although the approach based on T5 sentence embedding similarity and the *FLAN-T5-XXL* model have the same number of parameters (11B), the "traditional" cosine similarity based embedding classification seems to lag behind, indicating that prompt-tuning presumably learns a better representation of the task.

As we then increase the size of the underlying LLM even more, going from 11B to 62B parameters, we see yet another increase in performance, which is consistent with previous studies (e.g., (Chung et al., 2022; Mozes et al., 2023)) which have discussed scaling laws in more detail.

5. Case Study on Real Conversation Scenario

Our case study focuses on assessing the utility of ASR models for model deployment decisions on a moderately error-tolerant application: captioning person-to-person video conversations for speakers with impaired speech. Despite po-

Approach	approx. # params	full test set (940)	speech type prompted (391)	 unprompted (549)	severity severe (467)	 moderate (302)	 mild (149)
BERTScore + WER	350M	0.791	0.788	0.794	0.753	0.791	0.856
SentT5 Embedding Sim	11B	0.857	0.894	0.831	0.813	0.879	0.899
FLAN-T5 XXL	11B	0.878	0.903	0.860	0.836	0.923	0.890
LLM62B	62B	**0.900**	**0.918**	**0.886**	**0.863**	**0.944**	**0.903**

Table 4. AUC-ROC scores on full and subsets of the test set for the different approaches to predict meaning preservation of erroneous transcripts (numbers in brackets represent # examples in specific subset).

tentially higher error tolerance in this application scenario, minimum comprehensibility levels need to be taken into consideration to ensure that all users have a satisfactory experience.

An acceptable level of **Meaning Preservation Percentage** – the percent of examples where the transcription is either correct or if incorrect, still leads to preserved meaning – depends on each individual speaker and listener (e.g. familiar listener). However, based on anecdotal observations in this case study, we set this at 70% pending further investigation. The time required, following the process in Section 2, to label the semantic error severity of 100 utterances was about 1 hour , making this an expensive and hard to scale approach for deployment decisions. Using our LLM-based classifiers for meaning preservation, we propose the **LATTEScore (LLMs to Assess TranscripTion Errors Score)** as an estimate of the percentage of examples on which meaning was preserved:

$$\text{LATTEScore} = \frac{\text{\# Predicted Meaning Preserved}}{\text{\# Total Examples}} \times 100 \quad (1)$$

Our experiments below show that LATTEScore leads to better model deployment decisions than other metrics, like for example WER, which do not accurately convey the utility of transcripts.

5.1. Real Conversations data set

Because the domain that we tested was spontaneous and long form speech, the dataset we used comes from data collected in a "real conversation" setting from several speakers. Recordings were collected using a mobile speech recognition application. SLPs transcribed the recordings and confirmed that each was an extemporaneous utterance from a single speaker, in a free-form conversation, with no PII. Not only does this "real conversation" domain include rare words and named entities, but also, participants' speech was less articulated as compared to prompted speech recordings, and is a consequence of the conversational nature of the speech. We selected a subset of 10 speakers of varying types[4] and severities of speech impairment. The number of utterances in the real conversation data set varied from speaker to speaker. The total dataset consisted of 1031 utterances, out of which $\approx 80\%$ were incorrectly transcribed by personalized ASR models. See Table 5 for details.

Our original dataset used to train the meaning preservation classifiers (Section 3) contained some of the real conversation examples due to our original way of splitting by ground truth. Hence, we filtered out all utterances from both the training and the dev split which overlapped with the real conversation test set on ground truth. This leaves us with a new training set of 2558 utterances and a new dev set of 638 utterances. We retrain the $LLM62B$-based classifier, which was the best model according to Section 4, on this training set, pick the best checkpoint on this dev set, and report results on the real conversation test set.

5.2. Results

Overall, our $LLM62B$-based classifier achieves a ROC-AUC of 0.89 on the real conversation test set. We use the Recall-Precision curve on the dev set to identify the best decision threshold. Aimiming to avoid over-estimation of model quality (and hence false positive predictions for 'meaning preserved'), we chose a threshold of 0.85 which yields a very high precision of 90% (and an acceptable recall of 72%) on the dev set.

We can calculate the LATTEScore per speaker. Table 5 shows this score as well as the Percentage Meaning Preservation (i.e., the ground truth based on human assessment) and Word Accuracy (defined as $\textit{WordACC} = 100 - min(\textit{WER}, 100)$) per speaker. Comparing results for Percentage Meaning Preservation and LATTEScore, we see that in most cases we obtain a very similar assessment, with the exceptions being speakers $S3$ and $S6$, where LATTEScore under-estimates utility.

Looking at the *WordACC* scores, we can immediately see that those are not a good proxy for utility. For example

[4] Including Multiple sclerosis (MS), Amyotrophic lateral sclerosis (ALS), Primary lateral sclerosis (PLS), Vocal cord paralysis (VCP), Down Syndrome (DS)

Speaker	Etiology	Severity	Utterances	Word Accuracy	True Meaning Preservation Percentage	LATTEScore
S1	MS	Moderate	72	59.2	48.6	47.2
S2	Cleft Palate	Severe	94	60.0	35.1	34.0
S3	ALS	Severe	152	60.9	48.7	31.6
S4	PLS	Mild	61	66.7	44.3	55.7
S5	ALS	Severe	262	71.9	46.9	42.7
S6	VCP	Severe	50	72.6	**74.0**	64.0
S7	ALS	Moderate	179	**80.0**	52.0	52.0
S8	ALS	Moderate	76	**80.5**	57.9	55.3
S9	VCP	Severe	41	**86.5**	68.3	**70.7**
S10	DS	Mild	44	**90.8**	**77.3**	**77.3**

Table 5. Word Accuracy, Percentage Meaning Preservation (based on SLP assessment) and LATTEScore on real conversation test. Bolded numbers show which models would have been accepted based on our deployment decision thresholds (Word Accuracy $>=$ 80 and Meaning Preserved $>=$ 70).

the models of speakers $S7$ and $S8$ achieve high *WordACC* scores but their true utility is low. This is especially extreme for $S7$ where the Percentage Meaning Preservation is just above 50% (i.e., the meaning of every 2nd of the 179 examples of this speaker's test set is completely lost!) while *WordACC* $= 80\%$. SLPs reported that, for this speaker, the ASR model indeed performed very poorly in conversation.

For these experiments, we set the threshold for model acceptance based on **WordACC** to 80%, which is in line with what others have used before (e.g. (Tobin & Tomanek, 2022)). Based on feedback from a small sample of people using ASR, we set the threshold for Percentage Meaning Change and LATTEScore to 70%. However, note that an *acceptable* meaning preservation threshold is likely an individual decision made by the person using ASR, based on content, situation, conversation partner, amongst other factors. Table 5 shows that using LATTEScore leads to a much better deployment decision, very similar to that on the human assessment.

6. Conclusion and Future Work

Our results demonstrate that LLM-based classifiers can be used to reliably predict whether an erroneous ASR transcript can be considered meaning preserving or not, as compared to ground truth. This is especially helpful in scenarios where we expect high WERs, as is often the case for atypical speech, where conveying meaning appropriately is paramount. In our expermiments, LLM-based classifiers significantly outperformed our baselines that used syntax and sentence embeddings on this task. Further, the $62B$ parameters LLM outperformed a smaller LLM of $11B$ parameters. However, this performance comes at increased inference cost. Depending on the specific use-case, using the smaller, slightly worse performing model might be preferable.

Further, we have shown how LATTEScore can be used for model deployment decisions as it allows to estimate the overall percentage of meaning preservation by a model on a test set. We found that LATTEScore is better suited than WER (or equivalently Word Accuracy) to make such decisions. While it is tempting to apply a general threshold for required percentage of meaning preservation across all speakers, oftentimes one might have to decide on a case-by-case scenario depending on the user's specific context. E.g., a conversation with the doctor may require a higher threshold for meaning preservation compared to a casual chat with a familiar friend.

Ethical considerations. The work presented here focuses on disordered speech specifically, but its usefulness is not limited to only this population. The approach demonstrated here can be applied in any low resource scenario where human evaluation is beneficial but is less feasible due to time or cost constraints. It would be particularly interesting to study if it generalizes to errors observed with model transcripts from people of different ethnicities and demographics.

Future directions. Our LLM-based classifiers have been trained and tested on English examples only. In the future, we aim to study whether multi-lingual LLMs will allow for good zero-shot performance in other languages. Another extension of this work is to determine whether meaning preservation is achieved in a cross-sentence scenario: For some use-cases, eg conversations, overall meaning may still be preserved even if there are small meaning losses locally within individual sentences. We plan to leverage summarization techniques based on LLMs to assess this.

References

Semantic textual similarity on STS benchmark. https://paperswithcode.com/sota/semantic-textual-similarity-on-sts-benchmark. Accessed: 2023-03-03.

GPT4. https://openai.com/research/gpt-4. Accessed: 2023-05-24.

Cave, R. and Bloch, S. The use of speech recognition technology by people living with amyotrophic lateral sclerosis: a scoping review. *Disability and Rehabilitation: Assistive Technology*, pp. 1–13, 2021.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Dagenais, P. A., Adlington, L. M., and Evans, K. J. Intelligibility, comprehensibility, and acceptability of dysarthric speech by older and younger listeners. *Journal of Medical Speech-Language Pathology*, 19(4):37–49, 2011.

D'Innocenzo, J., Tjaden, K., and Greenman, G. Intelligibility in dysarthria: Effects of listener familiarity and speaking condition. *Clinical linguistics & phonetics*, 20 (9):659–675, 2006.

Green, J. R., MacDonald, R. L., Jiang, P.-P., Cattiau, J., Heywood, R., Cave, R., Seaver, K., Ladewig, M. A., Tobin, J., Brenner, M. P., Nelson, P. C., and Tomanek, K. Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases. In *Proc. Interspeech 2021*, pp. 4778–4782, 2021.

Hustad, K. C. The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. 2008.

Kim, S., Arora, A., Le, D., Yeh, C.-F., Fuegen, C., Kalinli, O., and Seltzer, M. L. Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. In *Proc. Interspeech 2021*, pp. 1977–1981, 2021. doi: 10.21437/Interspeech.2021-1929.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Mozes, M., Hoffmann, J., Tomanek, K., Kouate, M., Thain, N., Yuan, A., Bolukbasi, T., and Dixon, L. Towards agile text classifiers for everyone, 2023.

Ni, J., Ábrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., and Yang, Y. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J., and Woisard, V. Intelligibility and comprehensibility: A delphi consensus study. *International Journal of Language & Communication Disorders*, 57(1):21–41, 2022.

Tobin, J. and Tomanek, K. Personalized automatic speech recognition trained on small disordered speech datasets. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6637–6641, 2022. doi: 10.1109/ICASSP43922.2022.9747516.

Tobin, J., Li, Q., Venugopalan, S., Seaver, K., Cave, R., and Tomanek, K. Assessing ASR Model Quality on Disordered Speech using BERTScore. In *Proc. 1st Workshop on Speech for Social Good (S4SG)*, pp. 26–30, 2022. doi: 10.21437/S4SG.2022-6.

Yorkston, K. M., Strand, E. A., and Kennedy, M. R. Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology*, 5(1):55–66, 1996.

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021.

Zhou, K., Ethayarajh, K., Card, D., and Jurafsky, D. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 401–423, 2022.