LARGE LANGUAGE MODEL BASED GENERATIVE ERROR CORRECTION: A CHALLENGE AND BASELINES FOR SPEECH RECOGNITION, SPEAKER TAGGING, AND EMOTION RECOGNITION

Chao-Han Huck Yang[†1], *Taejin Park*[†1], *Yuan Gong*[†2], *Yuanchao Li*[†3], *Zhehuai Chen*[1], *Yen-Ting Lin*[4], *Chen Chen*[5], *Yuchen Hu*[5], *Kunal Dhawan*[1], *Piotr Żelasko*[1], *Chao Zhang*[6], *Yun-Nung Chen*[4], *Yu Tsao*[7], *Jagadeesh Balam*[1], *Boris Ginsburg*[1], *Sabato Marco Siniscalchi*[8], *Eng Siong Chng*[5], *Peter Bell*[3], *Catherine Lai*[3], *Shinji Watanabe*[9], *Andreas Stolcke*[10]

[1]NVIDIA [2]MIT CSAIL [3]University of Edinburgh
[4]National Taiwan University [5]Nanyang Technological University
[6]Tsinghua University [7]Academia Sinica [8]University of Palermo [9]CMU [10]Uniphore
{*hucky, taejinp*}*@nvidia.com* *yuangong@mit.edu* *yuanchao.li@ed.ac.uk* *andreas.stolcke@uniphore.com*

ABSTRACT

Given recent advances in generative AI technology, a key question is how large language models (LLMs) can enhance acoustic modeling tasks using text decoding results from a frozen, pretrained automatic speech recognition (ASR) model. To explore new capabilities in language modeling for speech processing, we introduce the generative speech transcription error correction (GenSEC) challenge. This challenge comprises three post-ASR language modeling tasks: (i) post-ASR transcription correction, (ii) speaker tagging, and (iii) emotion recognition. These tasks aim to emulate future LLM-based agents handling voice-based interfaces while remaining accessible to a broad audience by utilizing open pretrained language models or agent-based APIs. We also discuss insights from baseline evaluations, as well as lessons learned for designing future evaluations.

Index Terms— Language modeling, speech recognition postprocessing, speaker tagging, speech emotion recognition.

1. INTRODUCTION

Early statistical ASR systems based on the noisy channel model were conceived as utilizing two model components: acoustic model and language model (LM) [1]. First-pass decoding results could be subjected to postprocessing, or *rescoring*, to apply more powerful LMs or additional knowledge sources [2]. With the introduction of end-to-end (E2E) ASR systems in the early 2020s, language modeling for post-ASR has become more complex, e.g., by also modeling the implicit *internal LM* of E2E ASR systems [3–6]). With the advent of LLMs, however, post-ASR processing has become very attractive again, given the capacity of LLMs to model linguistic patterns, contextual influences, and even world knowledge as reflected in language.

More recently, LLMs and speech/language-model alignment methods have sparked considerable interest in new methods, such as cascaded LLM correction [7], for ASR and speech translation. LLM-based text-to-text generative error correction [7,8] has shown accuracy improvements over baseline rescoring methods, even surpassing n-best oracle performance, by bringing external knowledge to bear in ASR [9], speech translation [10], and image captioning [11, 12].

While a text-based ASR-LLM interface limits the richness of information utilized in post-processing, such as acoustic-prosodic expressions of speaker identity and paralinguistic properties, ASR outputs will still be sensitive to such information, especially when multiple hypotheses are output, effectively providing a text-based feature map that weakly reflects acoustic information [7].

Inspired by this observation, as well as by the early LLM-based studies cited, our challenge task aims to push research in two directions: (1) how large *performance gains* could be achieved by applying cascaded ASR-LLMs, and (2) how well cascaded LLMs could perform *tasks beyond word-level transcription*, such as recovering speaker and paralinguistic information. In other words, by leveraging LLMs, even text-only output from first-pass ASR system (such as available from a black-box API) might be enriched with paralinguistic and meta-information that is commonly thought to be encoded principally in acoustic-prosodic features. Fig. 1 illustrates the three challenge tasks, highlighting the assumption that ASR hypotheses in textual form contain sufficient implicit acoustic information to perform these tasks.

The short-term goal of the challenge is to introduce new ASR-LM tasks to the speech community that leverage the latest developments in LLM-based post-ASR text modeling, potentially benefiting the design of voice-interface LLM agents that use only text-based encodings. Through this initiative, we

[†]Equal contribution.

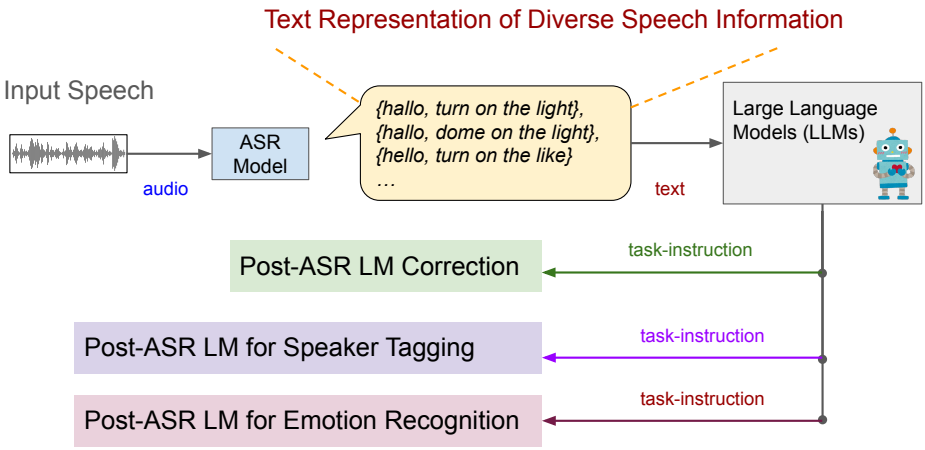

Fig. 1. The framework for LLM postprocessing of "text representation of speech" via ASR-decoded information, for three tasks: speech recognition, speaker diarization, and emotion recognition.

aim to advance the understanding of LLM capabilities in implicit acoustic modeling within the spoken language processing community. In the longer term, our goal is to highlight the importance of audio and other modalities for speech processing and understanding, setting the stage for future challenges that go beyond the initial text-only framework.

By promoting ongoing cooperation and knowledge sharing, we seek to catalyze significant progress in the field of LLM-based voice interfaces, driving innovation and opening new avenues for research, development, and practical applications of ASR-LLM. This includes error type analysis and cross-lingual variants of these tasks. In the following sections we give more background and introduce the specifics of the challenge tasks.

2. BACKGROUND

2.1. Post-ASR Text Modeling

This challenge considers an agent-based LLM application scenario with a fixed ASR interface. The focus is on characterizing how LLMs can enhance speech processing by leveraging the textual N-best hypotheses without explicit encoding of acoustic information. Additionally, we encourage participants to "push the limits of language modeling for ASR" given a setup based on a black-box ASR interface, which would be readily accessible through APIs, at modest cost.

The success of probabilistic language modeling systems in ASR can be traced back to several influential tools and frameworks, including SRILM [13], CNTK [14], and the LM components within Kaldi [15]. These tools have significantly contributed to the development and enhancement of speech recognition technology by providing robust methods for handling the complexities of speech processing. LLMs, on the other hand, have also benefited from democratized model inference (e.g., Claude) and open-source models (e.g., LLaMA) to establish end-to-end agent learning-based interfaces, such as AudioGPT [16]. For instance, work on task-activating prompting (TAP) [7] illustrates that instruction-prompted LLMs for ASR can correct recognition errors by inferring phonetic confusions or grammatical variants from the ASR-decoded text.

To investigate this form of ASR-LLM pipeline, we introduce three tasks based on ASR-decoded text, which have been studied previously and were shown to benefit from combined acoustic and language modeling. In this challenge, participants can explore a training-free setup by optimizing instruction prompts for the speech tasks, or by hosting LLMs in their own compute environment. To examine the limits of the text-only modality for speech processing, we limit the first SLT challenge to text-to-text modeling without access to acoustic embeddings. The acoustic information will be accessed through ASR hypotheses ranked by acoustic confidence scores and error word-level or utterance-level error attributions. We hope this simple setup will entice researchers without a speech background to become active in speech processing through language modeling.

2.2. Open Topics in LLM-based Speech Modeling

To avoid test set data that may have leaked into the pretrained LLMs, we prepare a non-public test set for each challenge subtask. While LLMs hold promise for post-ASR correction, they are not without problems. One concern is the potential for introducing biases reflected in the training data, which could affect the accuracy and fairness of the corrected transcripts. Additionally, LLMs could produce enriched text that diverges from the intended meaning or introduces new types of errors, necessitating novel error analysis methodologies. These potential issues highlight the need for ongoing research

and future challenges that assess the reliability and effectiveness of LLMs in ASR postprocessing.

Cross-modal setups will be incorporated into future versions of the challenge by providing acoustic embeddings [8] or raw waveforms. By connecting the latest research and developments in speech language modeling with practical applications, the challenge promotes implementation, adoption, and understanding of cutting-edge language technologies, including in scarce-training or low-compute scenarios. Participants can expect to foster the development of innovative solutions at the intersection of speech and language technology.

3. CHALLENGE DESCRIPTION

The GenSEC challenge at IEEE SLT 2024 consists of three tasks for post-ASR language modeling:

- **Task 1: Post-ASR Output Correction by LLM**

  The goal of this task is to map from N-best **H**ypotheses to ground **T**ruth transcriptions (H2T), similar to the setup in Yang et al. [7]. The training set includes recognition scores from various pretrained end-to-end ASR models and N-best hypotheses. Participants are allowed to use N-best hypotheses and their scores for re-ranking or generative correction to produce final transcriptions.

- **Task 2: Post-ASR Speaker Tagging Correction**

  This task aims at correcting the speaker tags in the output of a speaker-attributed (multi-speaker) ASR system. Speaker tagging in Task 2 refers to the speaker indices or anonymized speaker names (e.g., "speaker-A", "speaker-2") used to identify who spoke which words. We will provide errorful speaker-attributed transcripts produced by a multi-speaker ASR system. Participants in Task 2 are asked to submit corrected versions of the transcripts with accurate speaker tagging. A metric that gauges both speaker tagging and ASR accuracy will be used for evaluation. Similar to the other tasks, the current version of the Track-2 challenge allows use of the text modality only.

- **Task 3: Post-ASR Speech Emotion Recognition**

  This task aims to achieve utterance-level speech emotion recognition (SER) based on errorful ASR transcripts. Participants will develop ASR error correction methods combined with traditional deep-learning-based SER models, design novel prompt templates utilizing LLMs for SER, or utilize any other methods based on text input. Participants are encouraged to use the conversation as context to predict the emotion of a target utterance.

4. TASK DESCRIPTION

4.1. Task 1 on LLM for Post-ASR Output Correction

Background: Language model (LM) rescoring has been employed widely and for a variety of ASR technologies to improve ASR results, usually achieving good performance gains [2, 17–19]. In this approach, an external LM is trained separately and used to re-score the N-best hypotheses generated by the ASR system. While text error correction (TEC) has been explored [7, 20], ASR error correction is distinct due to the variability and distinct patterns of spoken language [21]. Neural models have been used widely with E2E models for text error correction or normalization [22–24]. These models often use beam search to generate new estimates, and can usually handle text normalization and denormalization of spelling errors.

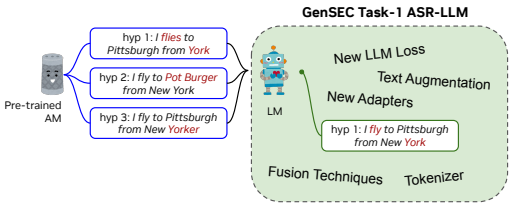

Fig. 2. Example Task 1 approach: post-speech recognition error correction with different techniques on LLMs.

Motivation: As shown in Fig. 2, with Task 1 we aim to explore the limits of ASR-LLM error correction, as well as how best to utilize the ambiguity conveyed by N-best output.

N-best dataset: The N-best open source corpus HyPoradise [9] will be made open-source under the MIT license. This includes HyPoradise training sets (316.8k pairs), development sets such as Librispeech-test-clean (2.6k pairs) and WSJ-dev93 (503 pairs), and evaluation sets including Librispeech-test-other (2.9k pairs) and WSJ-dev93 (333 pairs).

Baseline: We provide pretrained 1st-pass and 2nd-pass models. The details of existing engineering pipelines are listed below. Training code has been released[1] and the pretrained LLaMA2-7B model[2] has been released.

Evaluation: The challenge participants are allowed to apply their own 2nd-pass model to the provided ASR hypotheses decoded by beam search using Whisper.

The WER of the corrected hypotheses is used for evaluation. This WER is compared to two "oracle" WERs calculated from the N-best inputs, namely, 1) the lowest WER achievable by picking the best hypothesis from each N-best list, and 2) the compositional oracle method ocp: the achievable WER using "all tokens" in the N-best hypothesis list. The former can be viewed as a lower bound on re-ranking methods, while the latter denotes the lower bound using elements already occurring in the list. To understand the effect of text normalization (punctuation and capitalization, P&C) on ASR performance, both normalized and unnormalized (P&C) WERs are reported.

[1] https://github.com/Hypotheses-Paradise/Hypo2Trans

[2] https://huggingface.co/GenSEC-LLM.

Table 1. Task-1 WER (%) of post-ASR LM correction on the HyPoradise [9] dataset.

	train	**test**
Whisper-1.5B (first-pass) w/o LM	10.43	11.82
N-best Oracle	9.61	9.32
Reranking LM: T5-750M	9.90	9.74
Correction LM: T5-750M	9.21	9.05
Correction LM: LLaMA-13B	8.62	8.63
Correction LM: LLaMaA2-7B	8.71	8.33

Table 2. Task-1 dataset statistics: number of hypothesis-transcription pairs and average utterance length.

Domain							
Source	Category	Training Set	# Pairs	Length	Test Set	# Pairs	Length
LibriSpeech	Audiobooks	*train-960*	88,200	33.7	*test-clean* *test-other*	2,620 2,939	20.1 17.8
CHiME4	Noise	*train*	8,738	17.0	*test-real*	1,320	16.4
WSJ	Business news	*train-si284*	37,514	17.5	*dev93* *eval92*	503 333	16.7 17.3
SwitchBoard	Telephone	*train*	36,539	11.8	*eval2000*	2,000	11.8
CommonVoice	Accented English	*train-accent*	49,758	10.5	*test-accent*	2,000	10.5
Tedlium-3	TED talk	*train*	47,500	12.6	*test*	2,500	12.6
LRS2	BBC audio	*train*	42,940	7.6	*test*	2,259	7.6
ATIS	Airline info.	*train*	3,964	12.4	*test*	809	11.3
CORAAL	Interview	*train*	1,728	24.2	*test*	100	24.0
Total		*train*	316,881	18.1	*test*	17,383	14.1

4.2. Task 2: Post-ASR Speaker Tagging Correction

Background: While the use of lexical cues in speaker diarization, speaker turn detection, and speaker segmentation has been explored previously, it is still less commonly used than acoustic-only speaker diarization. Early studies in this area, such as those presented in [25, 26], utilized linguistic patterns to identify speakers during the diarization process.

Several studies have enhanced speaker segmentation and clustering accuracy by integrating ASR output to leverage lexical cues [27–29]. Furthermore, lexical cues can be incorporated into speaker diarization by combining speaker turn probabilities based on both audio and text during the clustering phase [30]. Alternatively, spoken words and speaker channels can be decoded jointly, thus utilizing lexical cues implicitly for diarization [31, 32].

More recently, the study presented in [33] introduced semantic information through neural embeddings generated by a spoken language processing (SLP) unit. Subsequently, a multimodal (audio-text) speaker change detector was proposed [34], along with a speaker error correction (SEC) system [35] based on a pretrained language model.

Due to the recent popularity of LLMs, the multi-speaker ASR and speaker diarization community has also begun employing LLMs to enhance performance. One framework established for this purpose fine-tunes PaLM 2-S [36] to correct speaker diarization errors from GCP's Universal Speech Model [37], which uses Turn-to-Diarize [28] for speaker diarization [38]. More recently, an ensemble of LLMs has been proposed to correct speaker diarization outputs [39].

Table 3. Task-2 cpWER (%) of the source files and the baseline system for text-only speaker recognition.

System	**dev**	**eval**
Source Transcript	24.65	28.45
Task-2 Baseline	24.54	28.37

Motivation: As discussed for Task 1, LM rescoring for ASR has been widely studied and adopted, as external language models can be trained on relatively larger text-only datasets. Numerous studies have demonstrated the benefits of using LLMs for speaker diarization correction. Despite an abundance of research, there has been no standardized evaluation of multi-speaker error correction systems. We believe that our GenSEC Challenge Task 2 is timely in filling this gap. To lower the bar for entry, we focus on the text modality, excluding acoustic or visual modalities in this first round. Therefore, in Task 2, we employ the system proposed in [40] without utilizing acoustic information from the (acoustic-only) speaker diarization system.

Datasets: The DiPCo [41], Mixer6 [42], AMI [43], and CallHome American English Speech (CHAES) [44] corpora have been divided into training, development, and evaluation sets. The session names have been anonymized to prevent participants from exploiting the publicly available ground-truth data. Additionally, the evaluation scripts display the total word count and the count of erroneous words, to verify whether the output transcripts have altered the total number of words. In total, there are 222 training samples, 13 development samples, and 11 evaluation samples. The dataset is accessible through Huggingface.[3]

Baseline: We generated the speaker-annotated transcripts from the system proposed in [45] as a baseline system,[4] based on NeMo [46] speaker diarization and NeMo ASR models. We provide the postprocessing method proposed in [40], where we replace the LLM with n-gram language models. Since the n-gram-based system shows similar results and has low computational demands, we use this n-gram baseline to gauge the performance of beam-search-based speaker tag correction. Fig. 3 shows how beam search decoding can correct the speaker tagging. To mask out acoustic information, the speaker probability values in Fig. 4 are all fixed at 0.96. Table 3 shows the accuracy of the baseline for development

[3] https://huggingface.co/datasets/GenSEC-LLM/SLT-Task2-Post-ASR-Speaker-Tagging

[4] https://github.com/tango4j/llm_speaker_tagging

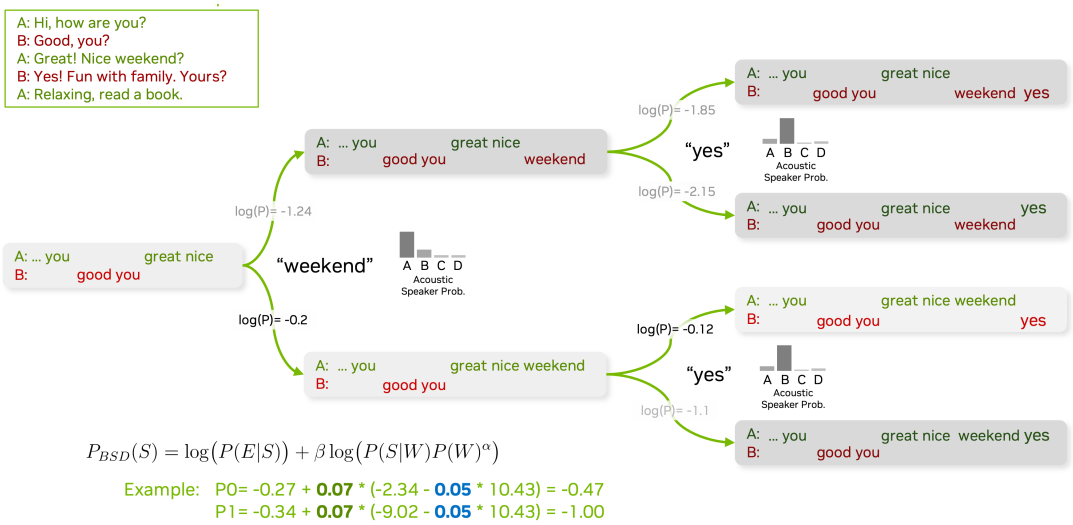

Fig. 3. Example Task 2 approach based on beam-search decoding for speaker tagging [40]

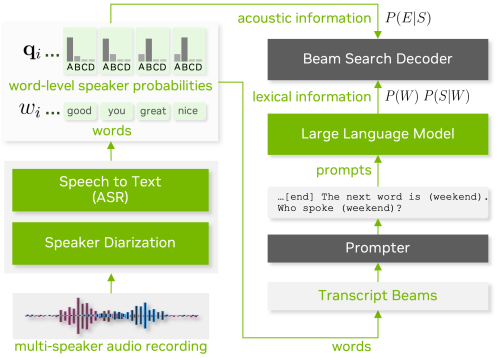

Fig. 4. Dataflow for the Task 2 baseline. Note that the acoustic-only diarization probability values are set to fixed values.

and evaluation sets.

Evaluation: We employ concatenated minimum permutation word error rate (cpWER), as presented in [47]. cpWER is calculated by concatenating the speaker-wise transcripts for every label permutation and selecting the permutation that results in the lowest WER, using the open-source and publicly available MeetEval [48] multi-speaker ASR evaluation toolkit. Additionally, we provide a Hugging Face-style leaderboard[5] for challenge participants to upload and evaluate their submissions.

4.3. Task 3: Post-ASR Speech Emotion Recognition

Background: Text-based SER has advanced significantly over the past decade. However, its use in real-world applications remains rare. One reason is that the majority of SER research relies on human annotation, i.e., manual transcripts. In contrast, even for elicited emotion corpora, transcripts from a state-of-the-art ASR system can result in high WERs [49], meaning that few findings obtained in the lab can be replicated in the wild. Moreover, SER on ASR transcripts is an understudied topic. Traditionally, researchers have considered confidence scores of recognized words [50], ASR error correction [51, 52], as well as fusion with audio information [53, 54] to mitigate the side effects of ASR errors. Still, there is a lack of comprehensive studies covering diverse situations (i.e., corpora, metrics, WERs, fusion techniques). With the rise of LLMs, it has become feasible to perform SER on ASR transcripts with simple prompting [55]. This emerging approach, however, has not been established as a reliable solution given the uneven performance with different prompting templates and the general lack of explainability of LLM outputs.

Motivation: Using text input only, without access to acoustic information, is a good starting point for exploring LLMs for SER, especially given that most LLMs are text-based. Insights gained about text-based SER, including handling of ASR errors, can be a foundation for future evaluations incorporating acoustic features. Potential future tasks could include ASR-error-robust multimodal fusion, enhancing word embeddings from ASR transcripts with discrete speech units, and developing ASR-integrated multimodal LLMs based on spoken language.

Dataset: We use the public IEMOCAP dataset [56]. Speech transcripts from eleven ASR models (Wav2vec2, HuBERT, Whisper, etc.) are provided for each audio segment [57]. We ask participants to predict four emotion classes: angry, happy (combined with excited), neutral, and sad, for each segment. All segments are presented in the order

[5] https://huggingface.co/spaces/GenSEC-LLM/task2_speaker_tagging_leaderboard

```
{
 "need_prediction": "yes",
 "emotion": "ang",
 "id": "Ses01F_script03_2_F027",
 "speaker": "Ses01_F",
 "groundtruth": "Oh shut up.",
 "hubertlarge": "oh shut u i will come",
 "w2v2100": "no shut a i wille tra",
 "w2v2960": "no shudder iwil e com",
 "w2v2960large": "o shudder obly come",
 "w2v2960largeself": "oh shudderco",
 "wavlmplus": "osabfa",
 "whisperbase": "Oh",
 "whisperlarge": "Oh shut up.",
 "whispermedium": "Oh shut up.",
 "whispersmall": "Oh shut up",
 "whispertiny": "Oh"
},
```


Fig. 5. Example of a data entry of Task 3.

of the conversation based on the timestamp. Participants are encouraged to use conversational context to improve emotion prediction.

An exemplary data entry is shown in Fig. 5, where *need_prediction* indicates whether this utterance should be included in the prediction procedure. "yes" denotes the utterances labeled with the four emotion classes and "no" denotes all other utterances. Note that we have removed the utterances that have no human annotations. The key *emotion* indicates the emotion label of the utterance. The key *id* indicates the utterance ID, which is also the name of the audio file in IEMOCAP dataset. The ID is exactly the same as the raw ID in IEMOCAP. The key *speaker* indicates the speaker of the utterance. The key *groundtruth* indicates the original human transcription provided by IEMOCAP while the remaining ten keys indicate the ASR transcription generated by the various ASR models.

Baseline: We provide two performance baselines with ASR transcripts from Whisper-tiny used as the text input: one with an LLM-based approach using GPT-3.5-turbo,[6] the other with a traditional approach based on a deep learning model. For the GPT-3.5-turbo approach, we performed zero-shot prediction with a context window of three (only previous utterances allowed), with code available[7] to participants as a reference. For the deep learning-based model, a two-layer feed-forward network was trained following the standard five-fold cross-validation of IEMOCAP. The first layer encodes RoBERTa output of dimension 768 into hidden states of dimension 128, and the second further encodes it into a dimension of 16. ReLU is used as the activation function between the layers. The dataset statistics and our baseline results are given in Table 4.

For the method based on GPT3.5-turbo, the accuracy is 44.70% on the training set and 55.18% on the test set. This

[6]Version: GPT-3.5-turbo-0125; Context window: 16,385 tokens; Training data: up to Sep 2021

[7]https://github.com/YuanGongND/llm_speech_emotion_challenge

Table 4. Task-3 dataset statistics and baseline unweighted accuracies (%).

	Training set	Test set
Number of samples (all)	5,525	4,730
Number of samples (four emotions)	2,577	2,923
Baseline accuracy (GPT3.5-turbo)	44.70	55.18
Baseline accuracy (traditional)	62.34	51.08

significant discrepancy is reasonable since the training set contains scripted dialogs (which may not align with emotion labels), while the test set is more spontaneous. For the traditional deep learning approach, the accuracy is 62.34% on the training set and 51.08% on the test set. This discrepancy is also plausible, considering duplicate textual scripts in the training set, which results in overlap between the training and development subsets of the training set.

Evaluation: We use unweighted four-class accuracy (number of correctly predicted samples / total number of samples). We will release a training set and a test set. Participants can use the training set to develop their methods and tune hyperparameters. The test set does not come with emotion labels or ground-truth transcription and is strictly disallowed for use in model development. We will rank the models based on their accuracy on the test set, but will also further evaluate the models with a separate unpublished test set to assess generalization. Participants are free to use any LLM (such as GPT or LLaMA) or non-LLM methods (traditional text-based emotion classifiers) based on the provided ASR transcriptions. For fairness considerations, participants should not use any audio data (including audio waveforms or acoustic features) or transcribe speech using their own ASR model. Participants are allowed to use additional training datasets, as long as they are specified and publicly available, but they must not include IEMOCAP. To encourage innovation, we do not place any other restrictions on the methods used, as long as they are automated.

5. CONCLUSION

We have created the GenSEC challenge to probe the capabilities of large language models for post-processing of ASR outputs. By standardizing the tasks, datasets and metrics, we hope to create a community of researchers that will advance the state of the art in speech processing systems by loose coupling of off-the-shelf ASR systems and techniques based on generative LMs, such as instruction prompting, text generation and in-context learning. We propose three tasks that go beyond speech transcription correction and include text-based speaker diarization correction and emotion recognition. Unlike traditional models, LLMs provide the potential for improving these tasks by leveraging linguistic and world knowledge learned during pretraining, and by taking advantage of long conversational context.

6. REFERENCES

[1] Frederick Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, 1976.

[2] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek, "Integration of diverse recognition methodologies through reevaluation of N-best sentence hypotheses," in *Workshop on Speech and Natural Language*, 1991.

[3] Zhong Meng, Sarangarajan Parthasarathy, Eric Sun, Yashesh Gaur, Naoyuki Kanda, Liang Lu, Xie Chen, Rui Zhao, Jinyu Li, and Yifan Gong, "Internal language model estimation for domain-adaptive end-to-end speech recognition," in *SLT*. IEEE, 2021.

[4] Mohammadreza Ghodsi, Xiaofeng Liu, James Apfel, Rodrigo Cabrera, and Eugene Weinstein, "Rnn-transducer with stateless prediction network," in *ICASSP*. IEEE, 2020.

[5] Yukun Ma, Chong Zhang, Qian Chen, Wen Wang, and Bin Ma, "Tuning large language model for speech recognition with mixed-scale re-tokenization," *IEEE Signal Processing Letters*, 2024.

[6] Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, et al., "Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition," in *ASRU*. IEEE, 2023, pp. 1–8.

[7] Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke, "Generative speech recognition error correction with large language models and task-activating prompting," in *ASRU*. IEEE, 2023.

[8] Srijith Radhakrishnan, Chao-Han Yang, Sumeer Khan, Rohit Kumar, Narsis Kiani, David Gomez-Cabrero, and Jesper Tegnér, "Whispering LLaMA: A cross-modal generative error correction framework for speech recognition," in *EMNLP*, 2023.

[9] Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng, "HyPoradise: An open baseline for generative speech recognition with large language models," *NeurIPS*, 2024.

[10] Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Dong Zhang, Zhehuai Chen, and Eng Siong Chng, "GenTranslate: Large language models are generative multilingual speech and machine translators," in *ACL*, 2024.

[11] David Chan, Austin Myers, Sudheendra Vijayanarasimhan, David Ross, and John Canny, "IC3: Image captioning by committee consensus," in *EMNLP*, 2023.

[12] Yusuke Hirota, Ryo Hachiuma, Chao-Han Huck Yang, and Yuta Nakashima, "From descriptive richness to bias: Unveiling the dark side of generative image caption enrichment," *arXiv preprint arXiv:2406.13912*, 2024.

[13] Andreas Stolcke et al., "SRILM-an extensible language modeling toolkit.," in *INTERSPEECH*, 2002.

[14] Frank Seide and Amit Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *KDD*, 2016.

[15] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *ASRU*. IEEE, 2011.

[16] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al., "AudioGPT: Understanding and generating speech, music, sound, and talking head," in *AAAI*, 2024.

[17] Andreas Stolcke, Barry Chen, Horacio Franco, Venkata Ramana Rao Gadde, Martin Graciarena, Mei-Yuh Hwang, Katrin Kirchhoff, Arindam Mandal, Nelson Morgan, Xin Lei, et al., "Recent innovations in speech-to-text transcription at SRI-ICSI-UW," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[18] Ebru Arisoy, Abhinav Sethy, Bhuvana Ramabhadran, and Stanley Chen, "Bidirectional recurrent neural network language models for automatic speech recognition," in *ICASSP*. IEEE, 2015.

[19] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[20] Daniel Dahlmeier and Hwee Tou Ng, "Better evaluation for grammatical error correction," in *NAACL-HLT*, 2012.

[21] Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik, "How might we create better benchmarks for speech recognition?," in *1st Workshop on Benchmarking: Past, Present and Future*, 2021.

[22] Hao Zhang, Richard Sproat, Axel H Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark, "Neural models of text normalization for speech applications," *Computational Linguistics*, 2019.

[23] Jinxi Guo, Tara N Sainath, and Ron J Weiss, "A spelling correction model for end-to-end speech recognition," in *ICASSP*. IEEE, 2019.

[24] Rao Ma, Mark JF Gales, Kate M Knill, and Mengjie Qian, "N-best t5: Robust asr error correction using multiple input hypotheses and constrained decoding space," *arXiv preprint arXiv:2303.00456*, 2023.

[25] Leonardo Canseco-Rodriguez, Lori Lamel, and Jean-Luc Gauvain, "Speaker diarization from speech transcripts," in *ICSLP*, 2004.

[26] Leonardo Canseco, Lori Lamel, and J-L Gauvain, "A comparative study using manual and automatic transcriptions for diarization," in *ASRU*. IEEE, 2005.

[27] Tae Jin Park and Panayiotis Georgiou, "Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks," in *INTERSPEECH*, 2018.

[28] Wei Xia, Han Lu, Quan Wang, Anshuman Tripathi, Yiling Huang, Ignacio Lopez Moreno, and Hasim Sak, "Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection," in *ICASSP*. IEEE, 2022.

[29] Aparna Khare, Eunjung Han, Yuguang Yang, and Andreas Stolcke, "ASR-aware end-to-end neural diarization," in *ICASSP*. IEEE, 2022, pp. 8092–8096.

[30] Tae Jin Park, Kyu J Han, Jing Huang, Xiaodong He, Bowen Zhou, Panayiotis Georgiou, and Shrikanth Narayanan, "Speaker diarization with lexical information," in *INTERSPEECH*, 2019.

[31] Laurent El Shafey, Hagen Soltau, and Izhak Shafran, "Joint speech recognition and speaker diarization via sequence transduction," in *INTERSPEECH*, 2019.

[32] Naoyuki Kanda, Xiong Xiao, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR," in *ICASSP*. IEEE, 2022, pp. 8082–8086.

[33] Luyao Cheng, Siqi Zheng, Zhang Qinglin, Hui Wang, Yafeng Chen, and Qian Chen, "Exploring speaker-related information in spoken language understanding for better speaker diarization," in *Findings of ACL 2023*, 2023.

[34] Jee-weon Jung et al., "Encoder-decoder multimodal speaker change detection," in *INTERSPEECH*, 2023.

[35] Rohit Paturi et al., "Lexical speaker error correction: Leveraging language models for speaker diarization error correction," in *INTERSPEECH*, 2021.

[36] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al., "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[37] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al., "Google USM: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[38] Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao, "DiarizationLM: Speaker diarization post-processing with large language models," *arXiv preprint arXiv:2401.03506*, 2024.

[39] Georgios Efstathiadis, Vijay Yadav, and Anzar Abbas, "LLM-based speaker diarization correction: A generalizable approach," *arXiv preprint arXiv:2406.04927*, 2024.

[40] Tae Jin Park, Kunal Dhawan, Nithin Koluguri, and Jagadeesh Balam, "Enhancing speaker diarization with large language models: A contextual beam search approach," *ICASSP*, 2023.

[41] Maarten Van Segbroeck, Ahmed Zaid, Ksenia Kutsenko, Cirenia Huerta, Tinh Nguyen, Xuewen Luo, Björn Hoffmeister, Jan Trmal, Maurizio Omologo, and Roland Maas, "DiPCo–Dinner Party corpus," *arXiv preprint arXiv:1909.13447*, 2019.

[42] Linda Brandschain, David Graff, Christopher Cieri, Kevin Walker, Chris Caruso, and A Neely, "The Mixer 6 corpus: Resources for cross-channel and text independent speaker recognition," in *Proc. of LREC*, 2010.

[43] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., "The AMI meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005.

[44] Alexandra Canavan, David Graff, and George Zipperlen, "CALLHOME American English speech," https://catalog.ldc.upenn.edu/LDC97S42, 1997.

[45] Tae Jin Park, He Huang, Ante Jukic, Kunal Dhawan, Krishna C Puvvada, Nithin Koluguri, Nikolay Karpov, Aleksandr Laptev, Jagadeesh Balam, and Boris Ginsburg, "The CHiME-7 challenge: System description and performance of nemo team's dasr system," *arXiv preprint arXiv:2310.12378*, 2023.

[46] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al., "NeMo: a toolkit for building AI applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.

[47] Shinji Watanabe, Michael Mandel, Jon Barker, et al., "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *CHiME Workshop*, 2020.

[48] Thilo von Neumann, Christoph Boeddeker, Marc Delcroix, and Reinhold Haeb-Umbach, "MeetEval: A meeting transcription evaluation toolkit," `https://github.com/fgnt/meeteval`, 2024.

[49] Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, and Catherine Lai, "ASR and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition," in *INTERSPEECH*, 2023.

[50] Jennifer Santoso, Takeshi Yamada, Shoji Makino, Kenkichi Ishizuka, and Takekatsu Hiramura, "Speech emotion recognition based on attention weight correction using word-level confidence measure.," in *INTERSPEECH*, 2021.

[51] Yuanchao Li, Pinzhen Chen, Peter Bell, and Catherine Lai, "Crossmodal ASR error correction with discrete speech units," in *SLT*. IEEE, 2024.

[52] Jiajun He, Xiaohan Shi, Xingfeng Li, and Tomoki Toda, "MF-AED-AEC: Speech emotion recognition by leveraging multimodal fusion, ASR error detection, and ASR error correction," *arXiv preprint arXiv:2401.13260*, 2024.

[53] Jeremy Ang, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg, and Andreas Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *INTERSPEECH*, 2002.

[54] Yuanchao Li, Peter Bell, and Catherine Lai, "Fusing ASR outputs in joint training for speech emotion recognition," in *ICASSP*. IEEE, 2022.

[55] Tiantian Feng and Shrikanth Narayanan, "Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting," in *ICASSP*. IEEE, 2024.

[56] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *LREC*, 2008.

[57] Yuanchao Li, Peter Bell, and Catherine Lai, "Speech emotion recognition with ASR transcripts: A comprehensive study on word error rate and fusion techniques," in *SLT*. IEEE, 2024.