FLANEC: EXPLORING FLAN-T5 FOR POST-ASR ERROR CORRECTION

Moreno La Quatra[⋆]*, Valerio Mario Salerno*[⋆]*, Yu Tsao*[†]*, Sabato Marco Siniscalchi*[‡]

[⋆]Kore University of Enna, [†]Academia Sinica, [‡]Università degli Studi di Palermo

ABSTRACT

In this paper, we present an encoder-decoder model leveraging Flan-T5 for post-Automatic Speech Recognition (ASR) Generative Speech Error Correction (GenSEC), and we refer to it as FlanEC. We explore its application within the GenSEC framework to enhance ASR outputs by mapping n-best hypotheses into a single output sentence. By utilizing n-best lists from ASR models, we aim to improve the linguistic correctness, accuracy, and grammaticality of final ASR transcriptions. Specifically, we investigate whether scaling the training data and incorporating diverse datasets can lead to significant improvements in post-ASR error correction. We evaluate FlanEC using the HyPoradise dataset, providing a comprehensive analysis of the model's effectiveness in this domain. Furthermore, we assess the proposed approach under different settings to evaluate model scalability and efficiency, offering valuable insights into the potential of instruction-tuned encoder-decoder models for this task.

***Index Terms*—** Flan-T5, Post-ASR Error Correction, Generative Error Correction, Encoder-Decoder Models

1. INTRODUCTION

Generative Speech Error Correction (GenSEC) focuses on correcting linguistic errors in transcriptions of spoken language. In Automatic Speech Recognition (ASR) systems, GenSEC acts as an effective post-processing step to enhance the accuracy [1, 2]. This step is crucial when ASR systems are used as inputs to downstream applications such as machine translation or information retrieval, where linguistic correctness is essential for maintaining the quality and reliability of the final output [3]. While modern end-to-end ASR systems have significantly improved transcription accuracy, their primary focus on acoustic modeling often results in neglecting grammatical error correction [4]. These systems excel at capturing acoustic cues and accurately transcribing spoken words. However, the inherent challenges of speech-to-text conversion, including variability in speech patterns, accents, and background noise, can lead to disfluencies, repetitions, and other speech irregularities that compromise the correctness of the final text. Therefore, ASR outputs may require additional post-processing to address linguistic errors and improve the overall quality of the transcription. The primary objective of GenSEC is to correct these errors by exploiting the context provided by n-best hypotheses generated by ASR models and leveraging external linguistic knowledge to produce a more accurate transcription. An n-best list, in this context, is an ordered list of the top n possible transcriptions generated by an ASR model, each with its associated confidence score. Recent advancements in natural language processing (NLP) have shown that large language models (LLMs) can be effectively adapted for various tasks. In this context, the Flan-T5 model [5] has demonstrated strong performance through instruction tuning, a process that specifically trains a single model to perform a diverse set of tasks by following specific instructions. This method significantly enhances the model's generalization capabilities, enabling it to adapt to specific tasks with minimal adjustments. By leveraging Flan-T5's performance in understanding and generating natural language [6], we aim to address the inherent challenges of post-ASR correction.

This paper presents our system for Task 1 of the GenSEC Grand Challenge, which focuses on post-ASR correction using large language models (LLMs). We propose FlanEC (*Flan*-T5 for *E*rror *C*orrection), an encoder-decoder model built on Flan-T5 [5]. Our contributions are as follows:

- We introduce FlanEC, an encoder-decoder model specifically instructed for GenSEC, trained at various scales from 250 million to 3 billion parameters.
- We evaluate FlanEC using the HyPoradise dataset [7], providing a comprehensive overview of its capabilities across different ASR domains.
- We explore the impact of scaling training data across all subsets of the HyPoradise dataset to assess whether training a single model on diverse domains can enhance performance across different subsets.
- We investigate the effectiveness of efficient model adaptation (i.e., LoRA [8]) versus full fine-tuning (FT) to assess to which extent it can provide comparable performance to full fine-tuning in our task.

2. POST-ASR ERROR CORRECTION

Correcting linguistic errors is a fundamental challenge in NLP and has traditionally focused on written text, with significant

advancements documented over the years [9]. LLMs have shown remarkable capabilities in understanding and generating natural language, making them well-suited for GenSEC tasks [10, 11, 5]. By integrating general knowledge and linguistic context, LLMs as a post-processing step can leverage the strengths of both ASR and NLP systems to enhance transcription accuracy.

2.1. Generative Speech Error Correction

GenSEC is an emerging field dedicated to enhancing the accuracy of ASR outputs by leveraging text-only information. It aims to generate a single correct transcription given one or more ASR hypotheses, each potentially accompanied by a confidence score. Previous works have demonstrated that integrating language models as an additional step on top of ASR models can significantly improve the quality of ASR transcriptions [1] with RNN-based LMs showing better performance than n-gram LMs [2]. Recent advancements in LLMs have further corroborated this hypothesis, showing that specific prompts, together with fine-tuning, can enhance the quality of the final transcriptions [12]. By combining the technical strengths of ASR systems with the advanced linguistic processing power of LLMs, GenSEC can lead to transcriptions that are more accurate and linguistically coherent.

In this context, the HyPoradise dataset [7] serves as a comprehensive benchmark for post-ASR error correction tasks, spanning several ASR domains to assess GenSEC models under various conditions. The dataset provides a standardized framework for assessing the effectiveness of different approaches to post-ASR error correction, emphasizing the linguistic aspects of the task without requiring additional acoustic modeling or alignment. Each spoken utterance within the dataset is accompanied by an n-best list of hypotheses generated by an ASR model (e.g., [4]), along with the corresponding reference transcription. Our work aims at exploring the potential of encoder-decoder language models to enhance GenSEC tasks using the HyPoradise dataset.

2.2. LLMs in GenSEC

Instruction tuning has highlighted the adaptability of LLMs, enabling them to effectively shift between different tasks by leveraging explicit instructions [13]. This approach enhances the model's ability to handle diverse NLP challenges efficiently. In-context learning [14, 15, 16] is a notable capability of LLMs, it allows models to learn directly from examples provided within a prompt, adapting to perform specific tasks using only the information given in the input. The emerging abilities of LLMs in this domain demonstrate their potential for handling a wide range of linguistic tasks [17]. In the context of GenSEC, LLMs are used to generate corrections for ASR hypotheses, simultaneously leveraging the context provided by the ASR outputs and the linguistic knowledge embedded in the model [18, 19]. Recent studies [20], however, have demonstrated that adapting model parameters outperforms in-context learning across various model scales. Thus, we focus on using an open-source model to tune parameters for post-ASR error correction.

While most LLMs are decoder-only, we focus on an encoder-decoder architecture that is open-source, available in various sizes, and capable of effectively following instructions. Flan-T5 [5] meets these criteria and has been pre-trained to solve 1,836 diverse tasks, significantly improving T5 [21] ability to generalize across different, even unseen, instructions. These characteristics make Flan-T5 particularly well-suited for GenSEC tasks, where the model must adapt to a range of linguistic contexts to generate accurate transcriptions by exploiting the context provided through ASR hypotheses. Flan-T5 has demonstrated superior performance in handling instruction-based tasks, even when compared to decoder-only LLMs of similar parameter scale, as evidenced by its results in the InstructEval benchmark [22].

3. FLANEC

The proposed FlanEC is an encoder-decoder model built on Flan-T5, specifically instructed for GenSEC tasks. We define a specific instruction template and use it to create FlanEC using HyPoradise dataset and leveraging the n-best lists generated by an ASR model.

3.1. Task prompt

The prompt for FlanEC is designed to instruct the model to generate a single correct transcription from a set of ASR hypotheses. To this end, the HyPoradise dataset, denoted as D, is used. It is composed of different subsets $D_1, D_2, \ldots, D_m$. A sample $s_{ij} \in D_i$, namely the j-th sample in the subset D_i, consists of (i) the n-best list of ASR hypotheses $H_{ij} = \{h_{ij1}, h_{ij2}, \ldots, h_{ijn}\}$, and (ii) the corresponding reference transcription r_{ij}. FlanEC is trained to map $H_{ij} \rightarrow r_{ij}$ by generating a single output written sentence that aligns with the reference transcription.

The task template is designed to be consistent with the original Flan-T5 approach, using natural language instructions without specific special tokens or tags [5]. The following prompt is intended to guide the model in generating the correct transcription from the n-best list:

Input: Generate the correct transcription for the following n-best list of ASR hypotheses:
- h_{ij1}
- h_{ij2}
- $\ldots$
- h_{ijn}

Output: r_{ij}

Where i denotes the subset index and j denotes the sample index within the subset. In our experiments, we use $n = 5$

Dataset	Training			Test		
	Avg. NT	% NT	% NS	Avg. NT	% NT	% NS
WSJ	0.83	3.19	40.74	0.46	1.86	25.00
ATIS	0.33	1.60	25.81	0.32	1.75	26.08
CHiME-4	2.28	9.83	61.06	0.93	4.21	47.65
Tedlium-3	0.49	2.26	42.80	0.30	1.07	34.11
CV-accent	1.40	8.47	51.36	1.46	8.74	51.90
SwitchBoard	1.32	6.04	71.08	1.24	5.80	70.95
LRS2	0.27	2.69	29.86	0.29	2.92	29.97
CORAAL	6.83	11.54	91.62	7.31	12.47	94.12
Overall	0.98	4.79	47.44	0.90	4.51	44.90

Table 1. Average and percentage of New Tokens (NT) present in reference transcriptions but absent in any of the n-best hypotheses, and the percentage of New Sentences (NS) where the reference does not match any n-best proposal, for training and test sets of each HyPoradise subset, and overall.

for the list of hypotheses, as [23] demonstrates this setting yields the best results. The model is trained to generate the correct transcription r_{ij} by exploiting the context provided in the list of hypotheses. To fully leverage the capabilities of Flan-T5, we did not apply any additional constraints on the decoding process, allowing the model to generate the output freely based on the input context. We speculate that a sufficient amount of training data, combined with the instruction-following capabilities of Flan-T5 and the context provided by the ASR hypotheses, should be enough to prevent the model from deviating from the desired output without limiting its ability to generate accurate transcriptions. Supporting this approach, our analysis presented in Table 1 shows that, on average across datasets, more than 4.5% of tokens in the reference transcriptions are not present in any of the n-best hypotheses. Notably, in the CV-Accent dataset, more than 8% of tokens in the reference transcriptions are new, underscoring the importance of allowing the model to generate beyond the provided hypotheses. Furthermore, the analysis reveals that a significant percentage of reference sentences do not match any of the sentences in the n-best lists. This indicates that limiting the decoding process to only the tokens or sentences that appear at least once in the original prompt, as suggested in [23], may restrict the model's correction capabilities and its use of internal linguistic knowledge.

3.2. Fine-tuning strategy

In line with Flan-T5, we evaluate two training settings:

- *SD*: Single Dataset, where we train a separate model on each subset of the HyPoradise dataset (e.g., ATIS, Tedlium3, etc.).
- *CD*: Cumulative Dataset, where we train a single model on the combined subsets of the HyPoradise dataset.

In *SD*, we train a separate model on each $D_i \in D$, whereas in *CD*, we train a single model on the union of all subsets, $\bigcup_{i=1}^{m} D_i$. While *SD* has been the standard approach for training GenSEC models [7], in *CD* settings, we aim to create an *omnibus* model capable of generalizing across different ASR domains. Our goal is to determine if FlanEC can generalize by leveraging multiple domains, similar to how Flan-T5 generalizes across different tasks. By combining diverse datasets such as ATIS (Airline Travel Information Systems), Tedlium3 (transcripts from TED talks), Switchboard (telephone conversation transcripts), and other training subsets in the HyPoradise dataset, we create a more varied training set. This approach utilizes the diversity of the training data to enhance the model's ability to correct errors across various ASR subsets. We hypothesize that training a single model on these diverse datasets will lead to better generalization and improved performance across multiple ASR domains. By learning from the combined data, the model can benefit from shared knowledge, potentially enhancing its overall performance.

To explore efficient fine-tuning methods, we also employ LoRA [8], which has shown promising results in adapting LLMs to specific tasks. LoRA involves training adapters on top of a pre-trained model, allowing for cost-effective fine-tuning without altering the original model's parameters. These adapters consist of additional, task-specific parameters trained to adapt the model to the target task. While LoRA is computationally efficient, it may not always match the performance of full fine-tuning (FT). We investigate this trade-off in the experimental section by training FlanEC using both LoRA and FT and comparatively evaluating their performance.

4. EXPERIMENTS

We evaluate FlanEC on the HyPoradise dataset, focusing on post-ASR error correction across different ASR domains. We adapt the model at various scales, ranging from 250 million to 3 billion parameters, to assess the impact of scaling model parameters on GenSEC performance. We compare the model's performance under different training settings, including *SD* and *CD*, to evaluate its generalization capabilities across diverse ASR domains. Additionally, we explore the effectiveness of LoRA versus full fine-tuning (FT) to assess the model's scalability and efficiency in the GenSEC task[1]. We refer to *adaptation* as the process of updating model parameters, whether through the LoRA method or full fine-tuning.

4.1. HyPoradise Dataset

Task 1 of the GenSEC Grand Challenge focuses on post-ASR error correction using the HyPoradise dataset [7]. This benchmark is designed to explore LLMs' capabilities in correcting

[1]Code to reproduce the experiments is available at: `https://github.com/MorenoLaQuatra/FlanEC`

ASR errors across various domains. It includes eight different subsets, each representing a distinct domain:

1. *WSJ* [24]: Business news and financial data readings.
2. *ATIS* [25]: Airline Travel Information Systems queries.
3. *CHiME-4* [26]: Noisy speech data from different environments, part of the CHiME-4 challenge.
4. *Tedlium-3* [27]: Transcripts from TED talks.
5. *CV-accent* [28]: Accented speech data from the English portion of the Common Voice dataset.
6. *SwitchBoard* [29]: Telephone conversation transcripts.
7. *LRS2* [30]: Audio clips from BBC programs.
8. *CORAAL* [31]: Interviews containing accented speech from regional varieties of African American English.

The diversity of the data ensures that models are evaluated across a wide range of ASR domains, providing a comprehensive assessment of their capabilities. Each subset contains both training and test data, comprising spoken utterances, n-best lists generated by an ASR model, and reference transcriptions. In our experiments, we train FlanEC-CD on $\bigcup_{i=1}^{m} D_i$ and evaluate its performance on the test subsets of the HyPoradise dataset. For FlanEC-SD, we train a separate model on each subset $D_i \in D$ and evaluate its performance on the corresponding test subset.

4.2. Experimental Setup

We train FlanEC in three different sizes: base (250M parameters), large (800M parameters), and extra-large (3B parameters). The models are fine-tuned using the pre-trained Flan-T5 checkpoint, with the number of parameters determined by the model size. We experiment with both LoRA and full fine-tuning (FT) to evaluate the model's performance under different training settings. In the case of LoRA, we use adapter layers for all fully connected layers available in the corresponding models. In all cases, we used the AdamW optimizer [32] and an effective batch size of 16.

When performing full fine-tuning, we use a maximum learning rate of 5×10^{-5}. For LoRA adaptation, we use a higher learning rate of 1×10^{-4} [8]. In both cases, we use a linear learning rate scheduler with a warm-up phase. Specifically, the learning rate is linearly increased from zero to the maximum value over the first 10% of the total training steps (warm-up phase). After the warm-up phase, the learning rate is linearly decreased back to zero for the remainder of the training. For FlanEC-CD, we set the maximum number of training epochs to 2, given the higher amount of training data and computational resources required for training on the combined dataset. For FlanEC-SD, we train each model for up to 10 epochs, following established best practices. In all cases, the best model is selected based on the validation WER and used for evaluation on the test subsets of the HyPoradise dataset. All experiments are conducted on 2 NVIDIA A100 GPUs with 80GB of memory each. We use the Hugging Face Transformers library [33] for model training and evaluation.

Evaluation Metrics: We evaluate the models using the Word Error Rate (WER) metric, which measures the percentage of words in the model's output that differ from the reference transcription. We follow the implementation provided in the original HyPoradise repository[2] to have consistent results.

5. RESULTS AND DISCUSSION

In this section, we present and analyze the performance of FlanEC on the HyPoradise dataset, focusing on post-ASR error correction across different ASR domains. The results are summarized in Table 2, where we compare the WER scores of FlanEC under different training settings, model sizes, and fine-tuning strategies. For an overall comparison, we have included the average WER across all datasets in the last row of Table 2. For a comprehensive evaluation, we also incorporate baseline results from the original HyPoradise [7] experiments, which are obtained using ASR without error correction. In addition, we evaluate two LLM-based approaches without fine-tuning: (1) Flan-T5 in a zero-shot (ZS) setting, using the 3B parameter model provided with the target sample and its 5-best hypotheses, and (2) in-context learning (ICL) using ChatGPT (`gpt-3.5-turbo-0125`), where the model also receives 5 training examples, each containing n-best lists and transcriptions. These training examples are selected based on the similarity of their first hypotheses to the target sample. We did not evaluate Flan-T5 in the ICL setting because prior research [34] has shown that encoder-decoder models struggle with ICL and even exhibit decreased performance when provided with in-context examples. These general-purpose LLM-based approaches, without task-specific adaptation, provide a broader context for evaluating post-ASR correction performance.

Single Dataset Models (SD): FlanEC-SD results indicate that scaling up the model size generally leads to improvements in WER. For instance, the 3B parameter model achieves the best performance on multiple datasets such as ATIS, CHiME-4, and CV-accent, with significant WER reductions compared to smaller models. The base model (250M parameters) trained with LoRA achieves better WER scores than the FT counterpart on the majority of subsets. This may be due to the base model's limited capacity, which can lead to overfitting when fine-tuning all model parameters using only a specific subset of data. For example, on SwitchBoard, the LoRA-adapted base model achieves a WER of 15.0%,

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

Dataset	Baseline	FlanT5 ZS	ChatGPT-3.5 ICL	FlanEC-SD						FlanEC-CD					
				LoRA			Finetuning			LoRA			Finetuning		
				250M	800M	3B	250M	800M	3B	250M	800M	3B	250M	800M	3B
WSJ	4.5	6.8	4.9	2.8	2.8	2.6	2.7	2.6	2.8	2.7	2.6	2.4	2.6	2.5	**2.3**
ATIS	8.3	18.6	3.7	1.3	3.2	1.4	1.7	1.5	**1.1**	2.7	3.0	2.1	2.3	1.3	**1.1**
CHiME-4	11.1	11.8	10.7	6.7	6.7	5.3	6.9	5.9	5.5	5.5	4.8	4.9	5.0	4.5	**3.9**
Tedlium-3	8.5	9.9	5.4	4.6	4.7	4.7	4.7	4.6	4.8	4.5	4.4	**4.0**	4.3	4.3	4.2
CV-accent	14.8	18.4	13.6	13.4	12.7	11.6	13.3	11.7	**10.8**	13.6	12.4	11.7	13.7	12.2	11.1
SwitchBoard	15.7	33.4	16.4	15.0	15.0	14.6	17.2	15.2	16.6	17.1	17.0	15.2	16.6	14.7	**14.1**
LRS2	10.1	42.3	11.0	9.1	9.0	8.6	9.3	**8.4**	**8.4**	10.0	9.3	9.5	10.6	9.6	8.7
CORAAL	**21.4**	25.7	24.8	31.2	23.5	23.5	26.6	22.2	23.5	23.2	22.5	22.9	22.9	22.4	22.2
Average	11.8	20.9	11.3	10.5	9.7	9.0	10.3	9.0	9.2	9.9	9.5	9.1	9.8	8.9	**8.5**

Table 2. Comparison of WER results for FlanEC-SD (Single Dataset, SD) and FlanEC-CD (Cumulative Dataset, CD). In the SD setting, each model is trained on a single subset of the HyPoradise dataset. In the CD setting, one model is trained on the combined subsets of the HyPoradise dataset. The models were adapted using both LoRA and full fine-tuning (FT) methods at different parameter scales. The best overall results are highlighted in **bold**. The last row reports the average across all datasets.

compared to 17.2% for the fully fine-tuned counterpart.

Cumulative Dataset Models (CD): The 3B FlanEC-CD model with full fine-tuning achieves the best overall performance in 4 out of 8 datasets, with significant WER reductions in domains such as WSJ, CHiME-4, and SwitchBoard. This indicates that certain domains may benefit from training on a cumulative dataset, leveraging the diverse linguistic patterns and contexts provided by multiple ASR domains. Consistent with SD results, full fine-tuning (FT) outperforms LoRA across all datasets. This is likely due to the scale of training data, which helps mitigate overfitting and allows the model to fully exploit its parameter capacity. While larger models generally deliver better performance, the 800M model often provides the best tradeoff between performance and computational complexity. It achieves substantial improvements over the base model while maintaining a lower computational overhead compared to the 3B model.

LoRA vs. Full Fine-tuning: Comparing LoRA and full fine-tuning (FT) strategies, FT consistently shows better performance across most datasets. The only exception occurs at smaller model sizes in the SD setting, where LoRA outperforms FT on the majority of the subsets, possibly because the limited model capacity leads to overfitting when all parameters are fine-tuned. LoRA mitigates this by adapting only a subset of parameters, reducing the risk of overfitting and task forgetting, especially in low-resource scenarios. However, most of the considered subsets contain around 4,000 samples or more, except for CORAAL, which has approximately 2,000 training examples. The results highlight that while LoRA is computationally efficient, GenSEC tasks significantly benefit from full model fine-tuning. In the CD settings, the 3B FlanEC FT model achieves an average improvement of 0.64 ± 0.46 absolute WER points compared to its LoRA counterpart. This highlights the advantages of full fine-tuning, especially as the scale of training datasets increases.

Model Size and Training Strategy: Our results clearly show that scaling up the model size and training on a combination of all training subsets with full fine-tuning can enhance the model's accuracy in post-ASR error correction. The 3B FlanEC-CD model stands out as the best-performing model across several subsets. The diverse nature of the HyPoradise dataset allows us to evaluate the generalization capabilities of FlanEC across various ASR domains. The results indicate that FlanEC-CD demonstrates robust generalization across different types of ASR data, from noisy environments (CHiME-4) to conversational speech (SwitchBoard).

However, consistent with previous findings [7], the model struggles with the CORAAL subset, with no model achieving a WER below the baseline. This aligns with existing research [35, 36, 37], which identifies acoustic modeling as the primary challenge for African-American language in ASR systems. As shown in Table 1, this dataset has the highest percentage of tokens in the reference transcriptions that are absent in any of the n-best hypotheses, with more than 90% of references lacking an exact match with any of the n-best hypotheses, likely contributing to its difficulty. Although larger models seem to mitigate the issue to some extent, the results suggest that further research is needed to address the challenges posed by this specific subset. Those findings emphasize the effectiveness of larger models and cumulative training datasets in enhancing post-ASR error correction, while also highlighting areas for improvement in handling highly challenging subsets like CORAAL.

Comparison with Additional Baselines: When evaluating LLM-based approaches without fine-tuning, we observe distinct patterns. First, Flan-T5 in the zero-shot (ZS) setting does not improve upon the baseline results. Although Flan-T5 is instruction-tuned, it struggles to generalize to the post-ASR correction task without any task-specific tuning, highlighting the limitations of using a model in a zero-shot configuration for this task. On the other hand, in-context learning (ICL)

with ChatGPT performs better than the baseline in 4 out of 8 cases, demonstrating the generalization capabilities of LLMs like ChatGPT, even when only provided with similar training examples rather than explicit fine-tuning. However, it is important to note that neither Flan-T5 (ZS) nor ChatGPT with ICL are specifically trained for this task; they are simply used as general-purpose LLMs applied to post-ASR correction. In contrast, FlanEC models, which are explicitly fine-tuned for the task, consistently outperform both Flan-T5 (ZS) and ChatGPT (ICL), demonstrating the need of task-specific adaptation to achieve optimal performance in this context.

6. CONCLUSIONS

We introduced FlanEC, an encoder-decoder model based on Flan-T5, for post-ASR error correction. Extensive experiments on the HyPoradise dataset demonstrate that FlanEC-CD, trained on a combined dataset, generally outperforms single dataset models, especially when fully fine-tuned. This suggests the potential for creating a general-purpose model capable of effective error correction across various ASR domains. Larger models showed significant performance gains, highlighting the importance of model scaling and diverse training data. While LoRA offers efficiency, full fine-tuning provides, on average, better results. Despite improvements, challenges with datasets like CORAAL remain, indicating areas for further research. We plan to expand our analysis to larger models (e.g., 11B parameters) to fully understand the impact of scaling and to further develop a versatile post-ASR correction model.

7. REFERENCES

[1] Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 369–375.

[2] Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5828.

[3] Sukhdeep S Sodhi, Ellie Ka-In Chio, Ambarish Jash, Santiago Ontañón, Ajit Apte, Ankit Kumar, Ayooluwakunmi Jeje, Dima Kuzmin, Harry Fung, Heng-Tze Cheng, et al., "Mondegreen: A post-processing solution to speech recognition error correction for voice search queries," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 3569–3575.

[4] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al., "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[6] Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang, "CoEdIT: Text editing by task-specific instruction tuning," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, Dec. 2023, pp. 5274–5291, Association for Computational Linguistics.

[7] 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," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[9] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant, "The CoNLL-2014 shared task on grammatical error correction," in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, Baltimore, Maryland, June 2014, pp. 1–14, Association for Computational Linguistics.

[10] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[12] 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 *2023*

IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2023, pp. 1–8.

[13] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi, "Cross-task generalization via natural language crowdsourcing instructions," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, May 2022, pp. 3470–3487, Association for Computational Linguistics.

[14] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui, "A survey for in-context learning," 2022.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[16] Siyin Wang, Chao-Han Yang, Ji Wu, and Chao Zhang, "Can whisper perform speech-based in-context learning?," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13421–13425.

[17] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus, "Emergent abilities of large language models," *Transactions on Machine Learning Research*, 2022, Survey Certification.

[18] Yuchen Hu, CHEN CHEN, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and EngSiong Chng, "Large language models are efficient learners of noise-robust speech recognition," in *The Twelfth International Conference on Learning Representations*, 2024.

[19] Kevin Everson, Yile Gu, Huck Yang, Prashanth Gurunath Shivakumar, Guan-Ting Lin, Jari Kolehmainen, Ivan Bulyko, Ankur Gandhe, Shalini Ghosh, Wael Hamza, Hung-Yi Lee, Anya Rastrow, and Andreas Stolcke, "Towards asr robust spoken language understanding through in-context learning with word confusion networks," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12856–12860.

[20] Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar, "Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, July 2023, pp. 12284–12314, Association for Computational Linguistics.

[21] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Google, Tech. Rep.*, 2019.

[22] Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria, "InstructEval: Towards holistic evaluation of instruction-tuned large language models," in *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, St. Julian's, Malta, Mar. 2024, pp. 35–64, Association for Computational Linguistics.

[23] Rao Ma, Mark J. F. Gales, Kate M. Knill, and Mengjie Qian, "N-best T5: Robust ASR Error Correction using Multiple Input Hypotheses and Constrained Decoding Space," in *Proc. INTERSPEECH 2023*, 2023, pp. 3267–3271.

[24] Douglas B. Paul and Janet M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[25] Charles T. Hemphill, John J. Godfrey, and George R. Doddington, "The ATIS spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990.

[26] Emmanuel Vincent, Shinji Watanabe, Jon Barker, and Ricard Marxer, "The 4th chime speech separation and recognition challenge," *URL: http://spandh. dcs. shef. ac. uk/chime challenge {Last Accessed on 1 August, 2018}*, 2016.

[27] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve, "Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*. Springer, 2018, pp. 198–208.

[28] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[29] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for re-

search and development," in *Acoustics, speech, and signal processing, ieee international conference on*. IEEE Computer Society, 1992, vol. 1, pp. 517–520.

[30] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6447–6456.

[31] Tyler Kendall and Charlie Farrington, "The corpus of regional african american language," *Version*, vol. 6, pp. 1, 2018.

[32] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.

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

[34] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. 2022, vol. 35, pp. 1950–1965, Curran Associates, Inc.

[35] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel, "Racial disparities in automated speech recognition," *Proceedings of the national academy of sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[36] Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew, "Uneven success: automatic speech recognition and ethnicity-related dialects," *Speech Communication*, vol. 140, pp. 50–70, 2022.

[37] Sean Robertson, Gerald Penn, and Ewan Dunbar, "Quantifying the role of textual predictability in automatic speech recognition," *arXiv preprint arXiv:2407.16537*, 2024.