Failing Forward: Improving Generative Error Correction for ASR with Synthetic Data and Retrieval Augmentation

Sreyan Ghosh[12*], **Mohammad Sadegh Rasooli**[3†], **Michael Levit**[2], **Peidong Wang**[2], **Jian Xue**[2], **Dinesh Manocha**[1], **Jinyu Li**[2],

[1]University of Maryland, College Park, USA [2]Microsoft, USA [3]Meta Reality Labs, USA

Correspondence: sreyang@umd.edu

Abstract

Generative Error Correction (GEC) has emerged as a powerful post-processing method to boost the performance of Automatic Speech Recognition (ASR) systems. In this paper, we first show that GEC models struggle to generalize beyond the specific types of errors encountered during training, limiting their ability to correct new, unseen errors at test time, particularly in out-of-domain (OOD) scenarios. This phenomenon amplifies with named entities (NEs), where, in addition to insufficient contextual information or knowledge about the NEs, novel NEs keep emerging. To address these issues, we propose **DARAG** (Data- and Retrieval-Augmented Generative Error Correction), a novel approach designed to improve GEC for ASR in in-domain (ID) and OOD scenarios. First, we augment the GEC training dataset with synthetic data generated using foundational generative models, thereby simulating additional errors from which the model can learn from. For out-of-domain scenarios, we simulate test-time errors from new domains similarly and in an unsupervised fashion. Additionally, to better handle NEs, we introduce retrieval-augmented correction wherein we augment the model input with entities retrieved from a datastore of NEs. Our approach is simple, scalable, and both domain- and language-agnostic. We experiment on multiple datasets and settings, showing that DARAG outperforms all our baselines, achieving 8%–30% relative WER improvements in ID and 10%–33% improvements in OOD settings. [1]

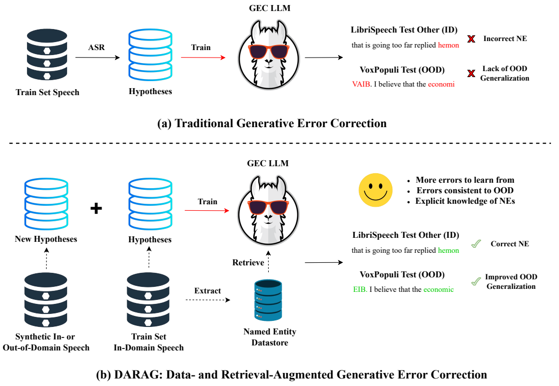
GEC LLM
ASR
Train
Train Set Speech
Hypotheses
LibriSpeech Test Other (ID)
that is going too far replied hemon
Incorrect NE
VoxPopuli Test (OOD)
VAIB. I believe that the economi
Lack of OOD Generalization
(a) Traditional Generative Error Correction
GEC LLM
+
Train
New Hypotheses
Hypotheses
More errors to learn from
Errors consistent to OOD
Explicit knowledge of NEs
LibriSpeech Test Other (ID)
that is going too far replied hemon
Correct NE
Retrieve
VoxPopuli Test (OOD)
EIB. I believe that the economic
Improved OOD Generalization
Extract
Synthetic In- or Out-of-Domain Speech
Train Set In-Domain Speech
Named Entity Datastore
(b) DARAG: Data- and Retrieval-Augmented Generative Error Correction

Figure 1: Comparison of traditional GEC and DARAG. We augment the training dataset with synthetic data generated using our algorithm and named entities retrieved from a datastore to improve in-domain and out-of-domain ASR.

1 Introduction

Automatic Speech Recognition (ASR) is the foundational task of converting spoken language into text. As a fundamental goal in computational language processing (Jurafsky, 2000), ASR has facilitated communication across diverse fields, including education (Caballero et al., 2017), healthcare (Latif et al., 2020), and others (den Bogaert et al., 2022). Advances in deep learning have driven significant progress in ASR, with end-to-end models achieving impressive results on various tasks (Li et al., 2022). However, one of the key challenges in real-world ASR applications (Li et al., 2015) is handling variations in speech due to factors like background noise (Chen et al., 2022), speaker accents (Turan et al., 2022), and different speaking styles (Syed et al., 2021). These factors lead to a significant reduction in the accuracy of ASR.

Humans demonstrate exceptional resilience to challenging speech conditions due to our inherent linguistic knowledge. Traditional ASR systems mimic this by incorporating a separate language model (LM) to rescore hypotheses during decoding (Toshniwal et al., 2018; Kannan et al., 2018). The LM evaluates the fluency of the N-best hypotheses generated by the ASR model, and the scores are combined with the ASR's own scores in a weighted fashion. The hypothesis with the highest combined score is then selected as the final transcript. However, the rise of large language models (LLMs) with advanced reasoning capabilities has

[*] Work done during an internship at Microsoft, USA

[†] Work done while at Microsoft

[1] For implementation details, please reach out to the author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 2466–2482
July 27 - August 1, 2025

opened possibilities beyond simple rescoring. This has led to the development of Generative Error Correction (GEC) (Chen et al., 2024), where models are trained to correct errors in the best hypothesis by leveraging information from other hypotheses, ultimately improving transcription accuracy.

GEC models are commonly trained on hypothesis-transcription pairs generated by ASR models using the training sets from a diverse range of ASR datasets. Recent approaches favor strong open-access ASR models for hypothesis generation (Chen et al., 2024; Hu et al., 2024a; Ghosh et al., 2024b) with the aim to generalize well across diverse datasets at test-time. In this paper, we investigate, for the first time, how the quality of training errors—specifically their nature, density, and distribution—impacts test-time performance across various settings. Through single-domain, single-dataset experiments (see Section 3), where GEC models are trained on the same datasets as their ASR counterparts, we observed minimal improvements in Word Error Rate (WER) for in-domain (ID) tests and no improvements for out-of-domain (OOD) tests. Upon closer examination, we attribute these shortcomings to three main factors:

1. ASR models generate too few errors on their training data for GEC models to effectively learn error correction.
2. GEC models are unable to generalize to the novel types of errors it sees at test time. This problem is exacerbated in OOD scenarios, where there is a significant shift in the nature of errors encountered during training versus those at test time.
3. GEC models continue to struggle with accurately correcting novel named entities (NEs) in transcriptions. While LLMs possess extensive linguistic knowledge, NEs often do not follow general language patterns. We attribute this challenge to insufficient context and a lack of knowledge about emerging NEs.

These observations lead us to a central hypothesis: *The generalization ability of GEC models is limited by the diversity and nature of error types encountered during training.* Improving performance requires training GEC models on a broader and diverse set of errors (for richer training signals) that are consistent in their characteristics with the types the ASR model generates on the test set. To better generalize to OOD, GEC models need to be trained to correct errors that the ID ASR model might plausibly make on the OOD test set.

Our Contributions. To this end, we propose **DARAG** (**Da**ta- and **R**etrieval-**A**ugmented Generative Error Correction), a simple, scalable, and domain-agnostic approach designed to boost GEC performance in ID and OOD scenarios. Our proposed approach is driven by the hypothesis that GEC models perform better when trained to correct errors they are likely to encounter at test time. To achieve this, DARAG generates domain-specific synthetic speech-transcript pairs using foundational generative models (LLMs and TTS models). The generated speech is then used to generate hypothesis-transcription pairs for training the GEC model. This process simulates errors that are specific to the target-domain vocabulary and also imitates the phonetic confusions that the ID ASR model would make in the target domain. Additionally, to improve named entity correction, inspired by RAG (Lewis et al., 2020), we introduce retrieval augmented correction (RAC). Specifically, we extract and store all named entities from the training dataset in a datastore and retrieve the top-k most similar entities during GEC. Our proposed method is scalable, with the datastore being easily extendable at test time to incorporate new entities as they are encountered. To summarize, our main contributions are as follows:

1. We conduct a first thorough investigation into the generalization limitations of LLM-based GEC, demonstrating that its performance can be improved by exposing it to diverse but consistent errors that ASR models are likely to produce at test time.
2. To address these challenges, we propose DARAG, a novel method for enhancing GEC in both ID and OOD scenarios. DARAG augments GEC training datasets with synthetic data and decouples named entity correction from the error correction learning process through RAG. DARAG significantly outperforms traditional GEC methods, improving ASR performance by 8%-33%.

2 Related Work

Generative Error Correction. Post-ASR error correction using language models (LMs) has been widely studied (Ma et al., 2023b,a; Zhang et al., 2023; Yang et al., 2023; Guo et al., 2019). Recently, large language models (LLMs) have been applied to this task, and the task has been known as generative error correction (Hu et al., 2024a; Ghosh et al., 2024b; Gu et al., 2024). While LLMs excel due to their advanced language comprehension, it remains unclear which errors they effectively correct, which they miss, and how well they handle unknown NEs that they lack prior knowledge of.

Domain Generalization and Named Entity in ASR. Transcribing NEs is a persistent challenge for ASR models (Das et al., 2022). Techniques such as memorization (Bekal et al., 2021) and biasing (Jayanthi et al., 2023) have been widely researched to improve NE transcription. However, these methods typically focus on known NEs seen during training and struggle with unseen entities, as autoregressive models tend to memorize NEs but generalize poorly to new ones (Heinzerling and Inui, 2020). Improving NE transcription using post-ASR processing or GEC has not been well explored. A parallel line of work also explores NER for ASR (Kumar et al., 2024; Yadav et al., 2020). ASR models often fail under distribution shifts, such as domain, accent, or dialect changes (Singhal et al., 2023). However, the robustness of GEC to domain shifts remains underexplored.

3 Preliminary

3.1 Problem Formulation

Let $\mathcal{D}^{\text{id}}_{\text{train}} = \{(a_i, t_i), 1 \leq i \leq n\}$ represent a human-annotated, in-domain dataset containing n pairs of speech and corresponding transcripts for training an ASR system ($\mathcal{D}^{\text{id}}_{\text{train}}$ is sourced from a single dataset and not pooled from multiple datasets unless otherwise mentioned). Consider $\mathcal{A}^{\theta}$ as an encoder-decoder ASR model trained on $\mathcal{D}_{\text{gold}}$. For GEC, our goal is to generate a list of N-best hypotheses h_i for each instance in $\mathcal{D}^{\text{id}}_{\text{train}}$ using beam search decoding. Next, using the hypotheses and corresponding gold transcripts, denoted by $\mathcal{H}^{\text{id}}_{\text{train}} = \{(h_i, t_i), 1 \leq i \leq n\}$, we fine-tune a language model to correct the errors in the best hypothesis by leveraging cues from the other N-1 hypotheses to directly produce an accurate transcription. During training, the true transcription t_i serves as the target. At inference time, for each instance in the test set $\mathcal{D}^{\text{id}}_{\text{test}}$, we generate a list of hypotheses and prompt the fine-tuned model to output a corrected transcript.

Test	ASR Train	Mismat. WER (↓)	Mat. WER (↓)
LS (Clean)	LS (960) (No GEC)	4.6	4.6
	LS (960)	4.4	4.4
	Vox	7.4	**3.9**
	SPGI	8.8	4.0
Vox	Vox (No GEC)	10.1	10.1
	Vox	9.4	9.4
	LS (960)	14.5	**6.9**
	SPGI	11.8	7.7
SPGI	SPGI (No GEC)	7.5	7.5
	SPGI	7.3	7.3
	LS (960)	14.2	**4.8**
	Vox	10.5	4.9

Table 1: Performance comparison of GEC across three different ASR benchmarks from three different domains. We evaluate and compare across two scenarios: (i) **Matched Scenario**: In this case, the hypotheses-transcription pairs for training our GEC model are derived from the Train split of the Test dataset (and not from the dataset the ASR model is trained on) (ii) **Mismatched Scenario**: In this case, the hypotheses-transcription pairs are derived from the same dataset the ASR model is trained on. We show that **(a)** For domain shifts, i.e., in cases where both the hypotheses and the ASR training dataset are from a domain different from the test, GEC leads to little to no improvement, and **(b)** For in-domain scenarios where only the hypotheses are derived from the same domain as the test, employing an ASR model trained on a different domain to derive the hypothesis boosts performance.

Our objective is to create a synthetic dataset, $\mathcal{D}^{\text{id}}_{\text{syn}} = \{(\hat{a}_i, \hat{t}_i), 1 \leq j \leq n_{\text{syn}}\}$, generate N-best hypotheses for each instance in it ($\hat{\mathcal{H}}^{\text{id}}_{\text{train}} = \{(\hat{h}_i, \hat{t}_i), 1 \leq j \leq n_{\text{syn}}\}$), and augment the original set $\mathcal{H}$ with $\hat{\mathcal{H}}$ to improve error correction on the test set $\mathcal{D}^{\text{id}}_{\text{test}}$. Alternatively, for an out-of-domain test set $\mathcal{D}^{\text{ood}}_{\text{test}}$, we assume the availability of a small train set from the same domain $\mathcal{D}^{\text{ood}}_{\text{train}} = \{(a_i, t_i), 1 \leq i \leq n_{\text{small}}\}$ where $n_{\text{small}} \ll n$ and the accompanying transcripts t_i may be human-annotated or generated from $\mathcal{A}^{\theta}$.

3.2 What do Error Correction Models Learn to Correct?

Most prior work on GEC models relies on foundational open-access ASR models, like Whisper, to generate hypotheses from various datasets and then trains GEC models on these hypotheses-transcription pairs, denoted as $\mathcal{H}^{\text{id}}_{\text{train}}$. However, because the training data used for such ASR models is often undisclosed, there is limited insight into the nature of errors present in the hypotheses and, consequently, the types of errors that the GEC models learn to correct. In this work, we aim to study error correction from a more transparent perspective. Table 1 presents experiments where we train an ASR model on a single dataset (LibriSpeech (LS) (Panayotov et al., 2015), VoxPopuli (Wang et al., 2021) (Vox), SPGIspeech (O'Neill et al., 2021)), then derive hypotheses from either the same or a different dataset, and use these pairs to train a GEC model. *This experimental setup* *proves to be more practical and reflective of real-world use-cases where users have the knowledge of errors and NEs learned during training and the test instances that are truely OOD.*Our key findings are as follows: **(i)** When GEC models are trained on a dataset in a different domain (i.e., both $\mathcal{D}^{\text{id}}_{\text{train}}$ and $\mathcal{H}^{\text{id}}_{\text{train}}$ come from a domain that is different from $\mathcal{D}^{\text{id}}_{\text{test}}$), no performance improvements are observed. We hypothesize this is due to the GEC model encountering errors at test time that differ significantly from those it saw during training. For instance, a hypothesis (HP)-transcription (GT) pair generated from the LibriSpeech train set using an ASR model trained on LibriSpeech is as follows:

> **GT:** biscuits with sugar on the top preserved ginger hams brawn under glass everything in fact that makes life worth living
> **HP 1:** biscuits with sugar on the top preserved ginger hams brawn under glass everything in fact that makes life worth living
> **HP 2:** biscuits with sugar on the top preserved ginger hams bran under glass everything in fact that makes life worth living

An error by the same ASR model on the VoxPopuli test set, is as follows:

> **GT:** spyware allows a third party to access the same data as the user.
> **HP 1:** spygware allows a third party to possess the same data as the user
> **HP 2:** spygware allows a third party to occupy the same data as the user

As we can see, it introduces semantic and lexical errors that are out of the domain knowledge learned during training. **(ii)** When GEC models are trained on a dataset in a similar domain (i.e., both $\mathcal{D}^{\text{id}}_{\text{train}}$ and $\mathcal{H}^{\text{id}}_{\text{train}}$ come from a domain identical to $\mathcal{D}^{\text{id}}_{\text{test}}$), improvements are minimal. We attribute this to the ASR model making fewer errors during inference, providing limited opportunities for the GEC model to learn effective corrections. For example, an ASR model trained on LibriSpeech and VoxPopuli have WERs of 2.2 and 5.1 on their respective train sets. **(iii)** To examine whether a higher error rate in hypotheses enhances GEC training, we use an ASR model trained on a different domain to generate hypotheses on our in-domain dataset $\mathcal{D}^{\text{id}}_{\text{train}}$ for GEC model training (the same ASR model is also used for test inference). Surprisingly, this setup consistently yields the most significant improvements, likely because the GEC model learns from a broader range of errors, enhancing its corrective abilities. *These findings highlight (i) the need for a large and diverse set of errors and (ii) the need for consistency in error characteristics with those that GEC models will encounter at test time.*

Test	**ASR Train**	**Mismat. F1** (↑)	**Mat. F1** (↑)
Vox	Vox (No GEC)	87.8	87.8
	Vox	87.8	87.8
	LS (960)	80.9	83.2
	SPGI	81.4	84.0

Table 2: Performance comparison of GEC on VoxPopuli, an entity-rich dataset. The Matched Scenario and Mismatched Scenarios are defined as in Table 1. We show that **(a)** For domain shifts, model performance degrades significantly on NEs. **(b)** For in-domain scenarios, GEC does not prove to be effective in correcting NEs.

3.3 How Well do they Fair on Named Entities?

To assess the ability of GEC models to correct named entities (NEs), we analyze their performance in various settings. As mentioned earlier, transcribing NEs is a major challenge in ASR, particularly in knowledge-rich domains. Table 2 compares GEC performance on VoxPopuli using models trained under different conditions. For this experiment, we leverage annotated NEs from the MSNER dataset (Meeus et al., 2024) for VoxPopuli. Our key findings are: **(i)** GEC models struggle to correct NEs, likely due to insufficient prior knowledge or context. **(ii)** In domain-shift scenarios, where ASR or GEC models have not encountered the target NEs during training, NE transcription accuracy declines sharply. *These results emphasize the importance of incorporating explicit knowledge of NEs to improve correction performance.*

4 Methodology

Fig. 2 illustrates our proposed method. We propose two simple extensions to improve conventional GEC. First, we propose training the GEC model on additional synthetic data generated using generative models. Additionally, instead of memorizing the named entities, we propose decoupling them from the learning process with RAG. To achieve this, we first extract named entities and store them in a datastore. During training and inference, we retrieve them from the datastore and augment them to the instruction with the best hypothesis and other hypotheses. In the following subsections, we explain our methodology in detail.

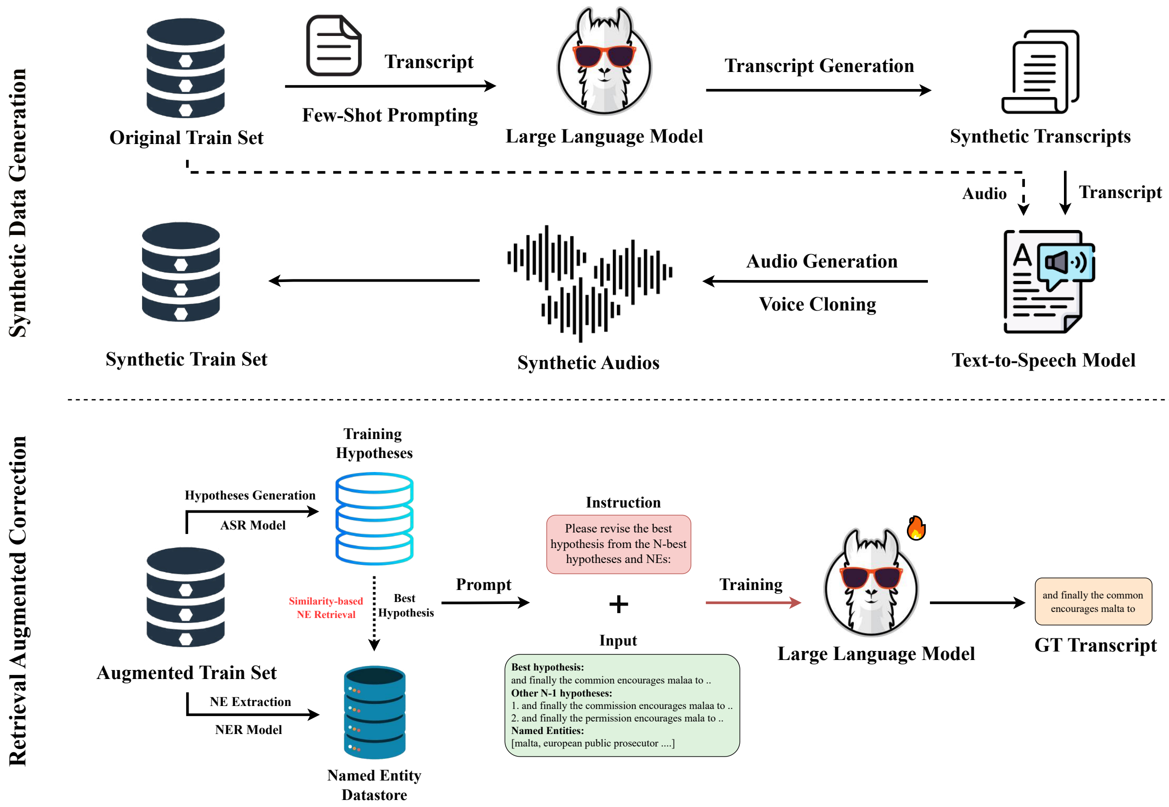

Figure 2: Illustration of **DARAG.** ① We generate synthetic data with LLMs and TTS models that are then used to generate hypotheses with diverse errors consistent with the types the ASR model generates on the test set. ② We extract the NEs and store them in a datastore. During training, for every instance, we retrieve the top-k most similar NEs to the best hypothesis and use it to construct an instruction-response pair. *Note that in OOD settings we only assume the availability of only a few unsupervised speech samples in the original train set, and pseudo-transcripts for prompting are generated using the in-domain ASR model.*

4.1 Synthetic Training Data Augmentation

For In-Domain Scenarios. As discussed in Section 3.2, GEC models fail to learn effective error correction due to the low number of errors in ASR training data. We hypothesize that generating novel spoken utterances not seen during ASR training will introduce more errors that can provide rich training signals for learning error correction. Our goal is to generate spoken utterances that closely mimic the speech characteristics of speakers in the same domain, replicating the style as if spoken by similar speakers in similar contexts. These utterances can then be used to generate new hypotheses, $\hat{\mathcal{H}}^{\text{id}}_{\text{train}}$, which we augment into the original dataset $\mathcal{H}^{\text{id}}_{\text{train}}$. We achieve this through a 3-step process:

Step 1. We prompt an LLM (LLaMa-2.0-Instruct (Touvron et al., 2023)) with in-context examples sampled from $\mathcal{D}^{\text{id}}_{\text{train}}$ to generate in-domain transcripts (prompt in Appendix B).

Step 2. Using voice cloning via TTS, we generate spoken utterances from the transcripts. The TTS model (Parler-TTS Mini (Lacombe et al., 2024)) is conditioned on randomly selected utterances from $\mathcal{D}^{\text{id}}_{\text{train}}$ to replicate the domain's speech style. Steps 1 and 2 ensure the generated utterances align with the domain and produce error patterns similar to those expected at test time.

Step 3. We generate hypotheses for these utterances using the ASR model $\mathcal{A}^{\theta}$. The resulting hypotheses, $\hat{\mathcal{H}}^{\text{id}}_{\text{train}}$, are then added to $\mathcal{H}^{\text{id}}_{\text{train}}$ to improve GEC model training.

For Out-of-Domain Scenarios. In OOD settings, we follow the same steps using $\mathcal{D}^{\text{ood}}_{\text{train}}$. If annotated transcripts are unavailable, we first transcribe the utterances with the ASR model $\mathcal{A}^{\theta}$. Recall that in our setting $\mathcal{D}^{\text{ood}}_{\text{train}}$ only has a few utterances ($n_{\text{small}} \leq 50$) and is unsuitable for adaptation of $\mathcal{A}^{\theta}$.

4.2 Retrieval Augmented Correction

To enhance the correction of NEs, we decouple NE correction from the main GEC process and introduce a Retrieval-Augmented Correction (RAC) approach (more in Appendix H). Inspired by RAG, we retrieve the most relevant NEs during both training and inference. Our method follows three steps:

Step 1. We apply NER on all transcriptions in the *train-set*, including those generated synthetically during the previous data augmentation step. We use SpaCy's en-core-web-sm model to extract all available NE types supported. The extracted NEs are stored in a datastore, $\mathcal{DS} = \{(s_t), 1 \leq t \leq d\}$ where d is the total number of extracted NEs.

Step 2. During GEC training and inference, we use SentenceBERT (Reimers, 2019) to retrieve the top-k NEs, $\overline{s}$, from $\mathcal{DS}$ based on their similarity to the best hypothesis (discussion on why SentenceBERT works can be found in Appendix F.). This is formally defined as:

$$\overline{s} = \text{top-}k_{1 \leq t \leq d}\left(\text{sim}\left(\frac{e_i \cdot e_t}{\|e_i\|\|e_t\|}\right)\right) \qquad (1)$$

where e_i is the SentenceBERT embedding for the best hypothesis, e_t is the embedding for an NE in $\mathcal{DS}$, and sim(.) is the cosine similarity between embeddings. We calculate similarity for each NE in $\mathcal{DS}$ and select the top-k most similar NEs. This simple method proves to be extremely effective in our case, as most errors in named entities belong to misspelled characters due to phonemes misrecognized by the ASR model. However, real-world datasets may contain multiple similarly spelled NEs, and retrieving all such NEs might make it difficult for error correction. We further discuss this in the limitations section.

Step 3. The retrieved NEs are then added to the input prompt during training and inference as a simple comma-separated list. We found that different prompt formats yielded similar results.

4.3 Fine-tuning

To train the LLM for error correction, we create instruction-response pairs and fine-tune our LLM on them. We employ the following template with the transcription as the target for fine-tuning:

> Below is the best hypothesis transcribed from a speech recognition system. Please try to revise it using the words that are only included in the other hypotheses and a list of named entities from a database, both of which will be provided to you.
> **Best-hypothesis**:
> **Other-hypothesis**:
> **Named-Entities**:
> **Response**:

Following prior work (Hu et al., 2021), we fine-tune only LoRA adapters.

5 Experimental Setup

Models and Hyper-Parameters. For our ASR model, we employ an encoder-decoder model with a 12-layer transformer-based encoder and a 6-layer conformer-based decoder. We train all datasets for 100 epochs with Adam optimizer, a learning rate of 1e-3, and an effective batch size of 128. For learning GEC, we train the LLaMa-2 $_{7\text{B}}$ (non-instruct) for 10 epochs with Adam optimizer, a learning rate of 5e-5, and an effective batch size of 32. We used a LoRA rank of 8, and we did not find a substantial change in performance by decreasing or increasing it. We generate n_{syn} = n or as many synthetic augmentations as the size of the original training set. For top-k NE retrieval, we set k=5. For N-best hypotheses, we set N=5. For OOD, we set n_{small}=100 and assume gold transcripts are not available. All results are averaged over 3 runs for 3 random seeds.

Datasets. We evaluated DARAG on 5 benchmark ASR datasets, including LibriSpeech-960 (LS), SPGISpeech (SPGI), VoxPopuli$_{\text{en}}$(Vox), Gigaspeech (Chen et al., 2021) (Giga) and TED-LIUM (Rousseau et al., 2012) (TED). Our OOD evaluation setup differs from prior works, and we explain our rationale in Appendix G.

Comparison Methods and Ablations. For comparison with DARAG, we employ (i) Baseline – Only ASR, and we perform no post-processing. (ii) Synth. Adap. – For ID, we add the synthetic data to the original ASR training data. For OOD, we do adapter-based continual fine-tuning of the ASR model (full-fine-tuning gave us worse performance) (iii) GER (Chen et al., 2024) – Vanilla GER, can also be considered as DARAG without data our retrieval augmentation (iv) RobustGER (Radford et al., 2023) (v) LM$_{rank}$ – We use the same LLM (continually fine-tuned on the text from training and synthetic dataset) as GER for re-scoring the N-best hypotheses and finally take the hypothesis with the best score averaged across the LLM and ASR model scores. (vi) Enhance – we also employ a speech enhancement front-end, a HiFi-GAN (Su et al., 2020), to denoise the noisy speech before passing it to the ASR model. For ablations, we employ (i) w/o RAC: DARAG without retrieval-augmented correction. (ii) w/o Aug.: DARAG without synthetic data augmentation but only retrieval augmentation based error correction. (iii) only Synth.: The GEC model is trained only on pairs from the synthetically generated data.

Table 3: Performance comparison (WER) of DARAG with other methods on various in-domain and out-of-domain settings (the Test is OOD w.r.t. the Train). We assume all 5 datasets are from different domains. We also report the absolute improvements w.r.t. the ASR-only Baseline. DARAG outperforms other methods by 8%–30% in in-domain and 10%–33% in OOD settings.

Test		Train	Baseline	Synth. Adapt.	+LM$_{rank}$	+Enhance	+GER	+RobustGER	+DARAG *(ours)*	w/o RAC *(ours)*	w/o Aug. *(ours)*	only Synth. *(ours)*
LS (Clean)	*In-Domain*	LS	4.6	4.6$^{+0\%}$	4.4$^{-4.3\%}$	4.5$^{-2.2\%}$	4.4$^{-4.3\%}$	4.4$^{-4.3\%}$	**4.0**$^{-13.0\%}$	4.2$^{-8.7\%}$	4.1$^{-10.9\%}$	4.6$^{+0.0\%}$
	Out-of-Domain	Vox	8.2	8.8$^{+7.3\%}$	8.1$^{-1.2\%}$	8.3$^{-1.2\%}$	7.4$^{-9.8\%}$	7.4$^{-9.8\%}$	6.1$^{-25.6\%}$	**5.9**$^{-28.0\%}$	6.8$^{-17.1\%}$	16.8$^{+44.8\%}$
		SPGI	8.9	9.0$^{+1.1\%}$	8.8$^{-1.1\%}$	8.8$^{-1.1\%}$	8.8$^{-1.1\%}$	8.6$^{-3.4\%}$	8.0$^{-10.1\%}$	**7.8**$^{-12.4\%}$	8.0$^{-10.1\%}$	
		TED	11.6	11.5$^{-0.9\%}$	11.1$^{-4.3\%}$	11.4$^{-4.3\%}$	11.3$^{-2.6\%}$	11.3$^{-2.6\%}$	10.2$^{-12.1\%}$	**9.9**$^{-14.7\%}$	10.9$^{-6.0\%}$	
LS (Other)	*In-Domain*	LS	8.4	8.3$^{-1.2\%}$	7.7$^{-8.3\%}$	7.2$^{-14.3\%}$	7.2$^{-14.3\%}$	6.9$^{-17.9\%}$	**6.4**$^{-23.8\%}$	7.0$^{-16.7\%}$	6.6$^{-21.4\%}$	8.0$^{-4.8\%}$
	Out-of-Domain	Vox	13.7	14.0$^{+2.2\%}$	13.5$^{-1.5\%}$	13.2$^{-1.5\%}$	13.5$^{-1.5\%}$	13.5$^{-1.5\%}$	**11.9**$^{-13.1\%}$	13.0$^{-5.1\%}$	13.0$^{-5.1\%}$	19.2$^{+7.2\%}$
		SPGI	14.2	15.5$^{+9.2\%}$	14.0$^{-1.4\%}$	13.5$^{-1.4\%}$	13.8$^{-2.8\%}$	13.8$^{-2.8\%}$	**12.6**$^{-11.3\%}$	13.4$^{-5.6\%}$	13.4$^{-5.6\%}$	
		TED	17.9	18.6$^{+3.9\%}$	17.9$^{+0.0\%}$	17.5$^{+0.0\%}$	17.4$^{-2.8\%}$	17.4$^{-2.8\%}$	**15.3**$^{-14.5\%}$	15.8$^{-11.7\%}$	16.0$^{-10.6\%}$	
Vox	*In-Domain*	Vox	10.1	9.9$^{-2.0\%}$	9.5$^{-5.9\%}$	9.9$^{-2.0\%}$	9.4$^{-6.9\%}$	9.4$^{-6.9\%}$	**8.6**$^{-14.9\%}$	9.4$^{-6.9\%}$	8.9$^{-11.9\%}$	9.5$^{-5.9\%}$
	Out-of-Domain	LS	14.9	15.2$^{+2.0\%}$	14.9$^{+0.0\%}$	14.9$^{+0.0\%}$	14.5$^{-2.7\%}$	14.5$^{-2.7\%}$	10.0$^{-32.9\%}$	**9.8**$^{-34.2\%}$	12.1$^{-18.8\%}$	19.8$^{+16.5\%}$
		SPGI	11.8	13.4$^{+13.6\%}$	11.4$^{-3.4\%}$	11.8$^{-3.4\%}$	11.8$^{+0.0\%}$	11.6$^{-1.7\%}$	**8.1**$^{-31.4\%}$	8.4$^{-28.8\%}$	10.3$^{-12.7\%}$	
		TED	17.0	18.6$^{+9.4\%}$	17.0$^{+0.0\%}$	17.2$^{+0.0\%}$	17.3$^{+1.8\%}$	17.3$^{+1.8\%}$	**14.4**$^{-15.3\%}$	14.7$^{-13.5\%}$	15.9$^{-6.5\%}$	
TED	*In-Domain*	TED	6.7	6.5$^{-3.0\%}$	6.6$^{-1.5\%}$	6.7$^{+0.0\%}$	6.6$^{-1.5\%}$	6.8$^{+1.5\%}$	**6.2**$^{-7.5\%}$	6.3$^{-6.0\%}$	6.6$^{-1.5\%}$	7.0$^{+4.5\%}$
	Out-of-Domain	SPGI	10.4	10.0$^{-3.8\%}$	10.2$^{-1.9\%}$	10.4$^{-1.9\%}$	10.8$^{+3.8\%}$	10.8$^{+3.8\%}$	8.8$^{-15.4\%}$	**8.1**$^{-22.1\%}$	10.1$^{-2.9\%}$	15.8$^{+51.9\%}$
		LS	9.1	9.0$^{-1.1\%}$	8.8$^{-3.3\%}$	9.1$^{-3.3\%}$	8.5$^{-6.6\%}$	8.5$^{-6.6\%}$	**8.2**$^{-9.9\%}$	8.7$^{-4.4\%}$	**8.2**$^{-9.9\%}$	
		Vox	9.9	10.8$^{+9.1\%}$	9.9$^{+0.0\%}$	9.9$^{+0.0\%}$	10.2$^{+3.0\%}$	10.2$^{+3.0\%}$	9.0$^{-9.1\%}$	**8.9**$^{-10.1\%}$	10.1$^{+2.0\%}$	
Giga	*In-Domain*	Giga	11.5	14.8$^{+28.7\%}$	10.8$^{-6.1\%}$	10.6$^{-7.8\%}$	11.0$^{-4.3\%}$	10.6$^{-7.8\%}$	**9.1**$^{-20.9\%}$	10.2$^{-11.3\%}$	9.5$^{-17.4\%}$	11.0$^{-4.3\%}$
	Out-of-Domain	TED	22.7	24.3$^{+7.0\%}$	21.5$^{-5.3\%}$	21.8$^{-5.3\%}$	22.3$^{-1.8\%}$	22.3$^{-1.8\%}$	**18.5**$^{-18.5\%}$	**18.5**$^{-18.5\%}$	21.3$^{-6.2\%}$	26.2$^{+15.4\%}$
		LS	18.0	23.4$^{+30.0\%}$	17.7$^{-1.7\%}$	17.5$^{-1.7\%}$	17.8$^{-1.1\%}$	17.8$^{-1.1\%}$	14.7$^{-18.3\%}$	**14.4**$^{-20.0\%}$	16.9$^{-6.1\%}$	
		Vox	16.3	20.2$^{+23.9\%}$	16.2$^{-0.6\%}$	16.2$^{-0.6\%}$	16.6$^{+1.8\%}$	16.6$^{+1.8\%}$	**14.5**$^{-11.0\%}$	15.0$^{-8.0\%}$	16.4$^{+0.6\%}$	
SPGI	*In-Domain*	SPGI	7.5	11.0$^{+46.7\%}$	7.1$^{-5.3\%}$	7.4$^{-1.3\%}$	7.3$^{-2.7\%}$	7.4$^{-1.3\%}$	**5.2**$^{-30.7\%}$	6.0$^{-20.0\%}$	6.4$^{-14.7\%}$	7.6$^{+1.3\%}$
	Out-of-Domain	TED	17.7	24.6$^{+39.0\%}$	17.4$^{-1.7\%}$	17.6$^{-1.7\%}$	17.7$^{+0.0\%}$	17.7$^{+0.0\%}$	**13.9**$^{-21.5\%}$	14.4$^{-18.6\%}$	17.0$^{-4.0\%}$	24.9$^{+40.7\%}$
		LS	14.4	18.1$^{+25.7\%}$	14.4$^{+0.0\%}$	14.4$^{+0.0\%}$	14.2$^{-1.4\%}$	14.2$^{-1.4\%}$	12.0$^{-16.7\%}$	**11.6**$^{-19.4\%}$	13.4$^{-6.9\%}$	
		Vox	11.3	14.7$^{+30.1\%}$	10.9$^{-3.5\%}$	11.0$^{-3.5\%}$	10.5$^{-7.1\%}$	10.4$^{-7.9\%}$	8.2$^{-27.4\%}$	**8.0**$^{-29.2\%}$	10.1$^{-10.6\%}$	

6 Results and Analysis

Main Results. Table 3 presents our main results, comparing performance across five datasets in both ID and OOD scenarios. *Our baseline results are analogous to those originally reported by ESPnet.* In the ID setting, the training and test sets come from the same dataset, whereas in the OOD setting, the training set is sourced from a different dataset, making the test set OOD for both the ASR and GEC models. For the OOD experiments, we randomly selected three datasets for training without any particular preference. Furthermore, we did not assume the availability of ground-truth transcripts in $\mathcal{D}^{\text{ood}}_{\text{train}}$ and instead used our ASR model to generate transcripts. Unlike previous experiments, we did not assume a separate dataset for ASR training; both the ASR model and the hypotheses were generated from the same training data. Our key findings can be summarized as follows: **(i)** DARAG substantially improves ASR performance for both ID (8%-30%) and OOD (10%-33%) settings. **(ii)** In ID settings, both RAC and synthetic augmentation prove essential, as ablating either component leads to decreased performance. **(iii)** In OOD settings, augmentation is more beneficial than RAC, likely because most NEs in the datastore do not match the NEs encountered during testing. **(iv)** DARAG proves to be a better way to use synthetic data to improve ASR as an adaptation with synthetic data leads to performance decrease over baseline.**(v)** In some OOD cases, removing RAC improves performance, which we attribute to mismatched OOD NEs, causing the GEC model to adjust certain NEs incorrectly. **(vi)** Relying solely on synthetic data is not effective for OOD scenarios, consistent with prior research indicating that human-annotated data remains crucial for optimal performance (Ghosh et al., 2024a). *Appendix C experiments show that DARAG does not replicate the original training data due to LLM memorization.*

6.1 Does Retrieval Augmentation Improve Transcription of Named Entities?

Table 4 presents a comparison of F1-micro scores for DARAG and various baselines in both ID and OOD settings. The results reveal several key insights: (i) DARAG consistently outperforms the baseline and conventional GEC approaches, with particularly large gains in OOD scenarios, demonstrating its robustness to domain shifts. (ii) Incorporating a datastore containing NEs from the in-domain dataset significantly improves OOD performance, in some cases matching the results of GEC models trained on ID datasets. This highlights the effectiveness of retrieval-augmented correction in enhancing ASR performance, including practical applications like meeting applications, where a

Test	Method	OOD F1 (↑)	ID F1 (↑)
Vox	Baseline	79.5	87.8
	+GEC	80.9	87.8
	+DARAG	82.3	90.0
	+synth. NE	82.8	**92.3**
	+DARAG w/ ID NE	89.9	-
	+synth. NE	**90.7**	-
LS (Other)	Baseline	82.5	93.2
	+GEC	82.0	93.5
	+DARAG	83.1	96.0
	+synth. NE	84.9	**96.4**
	+DARAG w/ ID NE	93.1	-
	+synth. NE	**93.4**	-

Table 4: Performance comparison of DARAG with other methods on the NE transcription. For ID, we employ the train set of the dataset as the test. For OOD, we employ LS for Vox and Vox for LS. w/ ID NE refers to DARAG, where the NE datastore is from the ID train set. w/ synth NE refers to additional synthetic NEs we add to the NE datastore.

datastore can be constructed with a list of relevant NEs and not necessarily included during training. (iii) Augmenting the datastore with synthetically generated NEs also shows promise in boosting DARAG's performance, indicating the potential to dynamically add emerging NEs to the datastore. This approach reduces the reliance on continual fine-tuning for ASR adaptation, which is typically required in other methods (Das et al., 2022).

6.2 DARAG for Source-Free UDA

Most Unsupervised Domain Adaptation (UDA) methods for ASR assume the presence of the entire unlabeled dataset from the target domain (Hu et al., 2024b). On the other hand, DARAG assumes the presence of only a few unlabeled instances. Fig. 3 shows DARAG proves to be effective for extreme low-resource UDA and outperforms STAR and continual fine-tuning with pseudo-labeling.

6.3 Real Data Outperforms Synthetic

Table 5 shows a comparison between DARAG and various baseline configurations where the synthetic dataset is replaced with the original training set of the target domain. The results clearly demonstrate that using real training data to generate GEC hypotheses significantly boosts performance, often surpassing complete ID settings. We attribute this improvement to two main factors: (i) the ASR model produces more errors on the GEC training dataset due to domain mismatch, providing richer training signals, and (ii) the datastore is enriched with real NEs from the original training set, offering more accurate context for corrections.

Extra Results. We present extra results in the Appendix, including ones for key hyper-parameter tuning, the importance of the voice cloning module, and the performance of open-access models like Whisper and Canary with DARAG. Additionally, we provide examples of generated augmentations in Table 13 and DARAG corrections in Table 14.

Test	Method	ASR Train	GEC Train	WER (↓)
Vox	Baseline	Vox	-	10.1
	+DARAG	Vox	Vox	8.6
	Baseline	LS	-	14.9
	Baseline	LS + Vox	-	10.3
	+DARAG	LS	LS	10.0
	+DARAG	LS	Vox	**6.9**
	Baseline	TED	-	17.0
	Baseline	TED + Vox	-	10.0
	+DARAG	TED	TED	14.4
	+DARAG	TED	Vox	7.5
SPGI	Baseline	SPGI	-	7.5
	+DARAG	SPGI	SPGI	5.2
	Baseline	LS	-	13.3
	Baseline	LS + SPGI	-	7.7
	+DARAG	LS	LS	12.0
	+DARAG	LS	SPGI	**4.8**
	Baseline	TED	-	17.7
	Baseline	TED + SPGI	-	7.9
	+DARAG	TED	TED	13.9
	+DARAG	TED	SPGI	5.0

Table 5: Performance comparison of DARAG in OOD settings with the baseline. We replace the generated augmentations with the original target domain training dataset (and do not generate extra augmentations). Training on hypotheses from the target domain train set leads to superior performance.

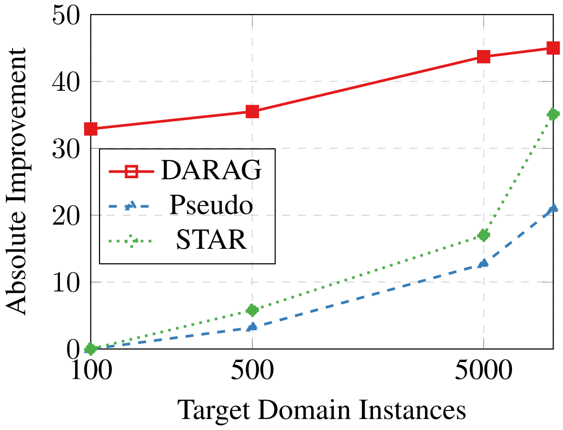

Figure 3: Comparison of DARAG with other methods on low-resource source-free UDA (LS → Vox). DARAG outperforms other methods with significant improvements.

7 Conclusion

We introduce DARAG, a novel approach to improve GEC for ASR. Our findings show that GEC models struggle to generalize in various ID and OOD cases. To address this, DARAG employs (i) synthetic data augmentation to simulate realistic test-time errors and (ii) retrieval-augmented NE correction. DARAG outperforms all compared methods, demonstrating its effectiveness.

Limitations

As part of future work, we would like to work on the following limitations of our proposed DARAG approach:

1. When the NE database is large, semantic similarity may result in the retrieval of multiple phonetically similar named entities, potentially causing confusion for the GEC model in choosing the correct entity. To address this, we plan to develop phoneme-aware NE retrieval methods to enhance retrieval accuracy.

2. The use of synthetic data generated by LLMs could introduce biases inherent to the language models, potentially affecting the GEC model's performance. In future work, we aim to explore strategies for mitigating such biases to ensure more robust error correction.

3. Although DARAG involves additional computational overhead for generating synthetic data, we anticipate that as model efficiency improves and lighter architectures become available, the overhead will be reduced, leading to even greater gains in performance. Additionally, our computational overhead is analogous to most prior synthetic data methods in speech (Gao et al., 2024a), vision (Azizi et al., 2023) or language (Ghosh et al., 2024c) and comparable to self-supervised learning, a well-known area of research for improving ASR performance.

4. We only study ASR datasets in the English language. Future work includes evaluating DARAG's performance in low-resource languages beyond English.

References

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. 2023. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*.

Dhanush Bekal, Ashish Shenoy, Monica Sunkara, Sravan Bodapati, and Katrin Kirchhoff. 2021. Remember the context! asr slot error correction through memorization. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 236–243. IEEE.

Daniela Caballero, Roberto Araya, Hanna Kronholm, Jouni Viiri, André Mansikkaniemi, Sami Lehesvuori, Tuomas Virtanen, and Mikko Kurimo. 2017. Asr in classroom today: Automatic visualization of conceptual network in science classrooms. In *Data Driven Approaches in Digital Education: 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia, September 12–15, 2017, Proceedings 12*, pages 541–544. Springer.

Chen Chen, Nana Hou, Yuchen Hu, Shashank Shirol, and Eng Siong Chng. 2022. Noise-robust speech recognition with 10 minutes unparalleled in-domain data. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4298–4302. IEEE.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. 2024. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36.

Guoguo Chen et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech 2021*.

Nilaksh Das, Monica Sunkara, Dhanush Bekal, Duen Horng Chau, Sravan Bodapati, and Katrin Kirchhoff. 2022. Listen, know and spell: Knowledge-infused subword modeling for improving asr performance of oov named entities. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7887–7891. IEEE.

Joachim Van den Bogaert, Laurens Meeus, Alina Kramchaninova, Arne Defauw, Sara Szoc, Frederic Everaert, Koen Van Winckel, Anna Bardadym, and Tom Vanallemeersch. 2022. Automatically extracting the semantic network out of public services to support cities becoming smart cities. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 343–344, Ghent, Belgium. European Association for Machine Translation.

Heting Gao, Kaizhi Qian, Junrui Ni, Chuang Gan, Mark A. Hasegawa-Johnson, Shiyu Chang, and Yang Zhang. 2024a. Speech self-supervised learning using diffusion model synthetic data. In *Forty-first International Conference on Machine Learning*.

Hongfu Gao, Feipeng Zhang, Wenyu Jiang, Jun Shu, Feng Zheng, and Hongxin Wei. 2024b. On the noise robustness of in-context learning for text generation. *arXiv preprint arXiv:2405.17264*.

Sreyan Ghosh, Sonal Kumar, Zhifeng Kong, Rafael Valle, Bryan Catanzaro, and Dinesh Manocha. 2024a. Synthio: Augmenting small-scale audio classification datasets with synthetic data. *arXiv preprint arXiv:2410.02056*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Purva Chiniya, Utkarsh Tyagi, Ramani Duraiswami, and

Dinesh Manocha. 2024b. LipGER: Visually-Conditioned Generative Error Correction for Robust Automatic Speech Recognition. In *Proc. INTERSPEECH 2024*.

Sreyan Ghosh, Utkarsh Tyagi, Sonal Kumar, Chandra Kiran Evuru, Ramaneswaran S, S Sakshi, and Dinesh Manocha. 2024c. ABEX: Data augmentation for low-resource NLU via expanding abstract descriptions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 726–748, Bangkok, Thailand. Association for Computational Linguistics.

Zijin Gu, Tatiana Likhomanenko, He Bai, Erik McDermott, Ronan Collobert, and Navdeep Jaitly. 2024. Denoising lm: Pushing the limits of error correction models for speech recognition. *arXiv preprint arXiv:2405.15216*.

Jinxi Guo, Tara N Sainath, and Ron J Weiss. 2019. A spelling correction model for end-to-end speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5651–5655. IEEE.

Benjamin Heinzerling and Kentaro Inui. 2020. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. *arXiv preprint arXiv:2008.09036*.

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

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and Eng Siong Chng. 2024a. Large language models are efficient learners of noise-robust speech recognition. In *International Conference on Learning Representations*.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Chengwei Qin, Pin-Yu Chen, Eng Siong Chng, and Chao Zhang. 2024b. Self-taught recognizer: Toward unsupervised adaptation for speech foundation models. *arXiv preprint arXiv:2405.14161*.

Sai Muralidhar Jayanthi, Devang Kulshreshtha, Saket Dingliwal, Srikanth Ronanki, and Sravan Bodapati. 2023. Retrieve and copy: Scaling asr personalization to large catalogs. *arXiv preprint arXiv:2311.08402*.

Daniel Jurafsky. 2000. Speech and language processing.

Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. 2018. 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)*, pages 1–5828. IEEE.

Rishabh Kumar, Sabyasachi Ghosh, and Ganesh Ramakrishnan. 2024. Beyond common words: Enhancing asr cross-lingual proper noun recognition using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6821–6828.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. 2024. Parler-tts. `https://github.com/huggingface/parler-tts`.

Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis. 2020. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jinyu Li, Li Deng, Reinhold Haeb-Umbach, and Yifan Gong. 2015. *Robust automatic speech recognition: a bridge to practical applications*. Academic Press.

Jinyu Li et al. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).

Danni Liu and Jan Niehues. 2024. Recent highlights in multilingual and multimodal speech translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 235–253.

Rao Ma, Mark JF Gales, Kate M Knill, and Mengjie Qian. 2023a. N-best T5: Robust ASR error correction using multiple input hypotheses and constrained decoding space. *arXiv preprint arXiv:2303.00456*.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill. 2023b. Can generative large language models perform asr error correction? *arXiv preprint arXiv:2307.04172*.

Quentin Meeus, Marie-Francine Moens, and Hugo Van hamme. 2024. MSNER: A multilingual speech dataset for named entity recognition. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 8–16, Torino, Italia. ELRA and ICCL.

Patrick K O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D Shulman, et al. 2021. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv preprint arXiv:2104.02014*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015*

IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5206–5210. IEEE.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, et al. 2024. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv preprint arXiv:2401.16658*.

Krishna C Puvvada, Piotr Żelasko, He Huang, Oleksii Hrinchuk, Nithin Rao Koluguri, Kunal Dhawan, Somshubra Majumdar, Elena Rastorgueva, Zhehuai Chen, Vitaly Lavrukhin, et al. 2024. Less is more: Accurate speech recognition & translation without web-scale data. *arXiv preprint arXiv:2406.19674*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.

Ashish Seth, Sreyan Ghosh, S. Umesh, and Dinesh Manocha. 2024. Stable distillation: Regularizing continued pre-training for low-resource automatic speech recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10821–10825.

Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2023. Domain adaptation: challenges, methods, datasets, and applications. *IEEE access*, 11:6973–7020.

Jiaqi Su, Zeyu Jin, and Adam Finkelstein. 2020. Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. *arXiv preprint arXiv:2006.05694*.

Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating informative conclusions for argumentative texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3482–3493, Online. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tugtekin Turan, Dietrich Klakow, Emmanuel Vincent, and Denis Jouvet. 2022. Adapting language models when training on privacy-transformed data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4367–4373, Marseille, France. European Language Resources Association.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. *arXiv preprint arXiv:2005.11184*.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Ziji Zhang, Zhehui Wang, Rajesh Kamma, Sharanya Eswaran, and Narayanan Sadagopan. 2023. Patcorrect: Non-autoregressive phoneme-augmented transformer for asr error correction. *arXiv preprint arXiv:2302.05040*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, et al. 2023. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, pages 57–68.

A Appendix

In the Appendix, we provide:

1. Section B: Prompts

2. Section C: Analysis of LLMs copying the training data
3. Section D: Performance of DARAG w/o Voice Cloning
4. Section E: Results of DARAG with Foundational ASR Models
5. Section F: Why is SentenceBERT an effective NE Retriever?
6. Section G: Rationale behind our OOD setup
7. Section H: Why generating synthetic transcripts from noisy hypothesis proves to be effective?
8. Section I: In-Domain Performance After Out-of-Domain Adaptation
9. Section J: Key Hyper-Parameter Tuning Results
10. Section K: Examples of DARAG Generations

B Prompts

We prompt LLaMa-2-Instruct in batched mode with a temperature of 0.7 and top-p of 1. We use this setting throughout all our experiments for generation and correction. We use the below prompt to generate synthetic transcripts using LLaMa-2-Instruct:

> You need to act as a synthetic data generator. I will provide you with some example transcripts from a speech recognition dataset that I have transcribed using an ASR model. The transcripts are not related to each other. You need to first understand the nature of the spoken utterances from the transcripts and analyze their distinct features, like domain, style, length, etc. Next, with what you understood, you need to generate 2 short and diverse utterances with the same properties but diverse content. Each utterance should be a single sentence. Please include named entities as and when possible, but it is not necessary. Keep the utterances short and in line with the examples. Your generated transcripts should be coherent. Here are the example transcripts, one in each line:{ }. Return a JSON with 2 keys named "First Transcript" and "Second Transcript" with the values as the generated transcripts.

C Are LLMs Just Replicating the Original Training Data?

Previous research has suggested that LLMs may memorize open-domain ASR training transcripts (Liu and Niehues, 2024; Team et al., 2023), raising the risk of replicating training data while generating synthetic data. To evaluate whether this occurs with DARAG, we perform two checks: (i) We use SentenceBERT to calculate the cosine similarity between each generated transcript and all transcripts in the original training set, reporting the average semantic similarity across instances in Table 6 ii) We compute the BLEU score for each generated transcript, using the transcript with the highest cosine similarity from the previous step as a reference. Table 6 shows the average BLEU scores across BLEU$_1$, BLEU$_2$, and BLEU$_3$. The low BLEU scores indicate that DARAG does not simply replicate the training data. The semantic similarity indicates that DARAG generates transcripts that are consistent with the domain.

Dataset	Similarity	BLEU
LS	0.32	0.12
Vox	0.29	0.10
SPGI	0.25	0.06
Giga	0.22	0.13
TED	0.26	0.14

Table 6: Semantic similarity and BLEU scores between original and generated transcripts across all datasets.

D DARAG w/o Voice Cloning

Table 7 compares the performance of DARAG in both ID and OOD scenarios, with and without voice cloning. As discussed in Section 4.1, voice cloning via TTS allows the model to generate synthetic speech that, when transcribed, produces hypotheses containing errors similar to those encountered during testing in that domain. As shown in the table, DARAG experiences a performance drop without voice cloning, with a more significant decline in OOD scenarios.

E Results of DARAG with Foundational ASR Models

Table 8 compares the performance of DARAG on 5 datasets for 3 foundational ASR models, Whisper $_{\text{Large-v3}}$, OWSM (Peng et al., 2024), $_{\text{Large}}$ and Canary (Puvvada et al., 2024). For ID settings

Test	Method	Train	WER (↓)
Vox	Baseline	Vox	10.1
	+DARAG	Vox	8.6
	+DARAG w/o Voice Cloning	Vox	8.8
	Baseline	LS	14.9
	+DARAG	LS	10.0
	+DARAG w/o Voice Cloning	LS	12.2
LS (Other)	Baseline	LS	8.4
	+DARAG	LS	6.4
	+DARAG w/o Voice Cloning	LS	7.3
	Baseline	Vox	13.7
	+DARAG	Vox	11.9
	+DARAG w/o Voice Cloning	Vox	14.5

Table 7: Performance comparison of DARAG with and without voice cloning. Performance drops sharply without voice cloning, especially in OOD scenrios, thereby confirming the importance of the voice cloning for generating augmentations.

where the ASR model is already trained on one of the datasets, we adhere to our ID experimental setup as mentioned in Section 4. For OOD, we adhere to our OOD experimental setup as mentioned in Section 4. As we can clearly see, DARAG improves the performance of foundational ASR models by significant margins, thereby showing promise in applications with foundational ASR models trained on multiple datasets.

	LS Clean	VOX	TED	GIGA	SPGI
Whisper $_{\text{Large}}$	2.0$^{\text{OOD}}$	9.8$^{\text{OOD}}$	3.9$^{\text{OOD}}$	10.4$^{\text{OOD}}$	3.0$^{\text{OOD}}$
Whisper $_{\text{Large}}$ + DARAG	1.9$^{\text{ID}}$	9.2$^{\text{OOD}}$	3.4$^{\text{OOD}}$	10.0$^{\text{OOD}}$	2.7$^{\text{OOD}}$
OWSM	2.7$^{\text{ID}}$	7.2$^{\text{ID}}$	4.8$^{\text{ID}}$	11.2$^{\text{OOD}}$	-
OWSM + DARAG	2.4$^{\text{ID}}$	6.9$^{\text{ID}}$	4.3$^{\text{ID}}$	10.5$^{\text{OOD}}$	-
Canary	1.9$^{\text{ID}}$	5.8$^{\text{ID}}$	3.6$^{\text{OOD}}$	10.1$^{\text{OOD}}$	2.1$^{\text{OOD}}$
Canary + DARAG	1.8$^{\text{ID}}$	5.5$^{\text{ID}}$	3.2$^{\text{OOD}}$	9.8$^{\text{OOD}}$	1.9$^{\text{OOD}}$

Table 8: Results for DARAG when coupled with foundational ASR models. For a model already trained on a respective dataset, we label it with ID and OOD otherwise.

Why is the comparison not made in the main paper? We do not compare with foundational open-access models like Whisper as it contradicts the primary motivation of our work. Such models do not disclose the datasets used for training, making it impossible to determine which datasets are ID and which are out-OOD. Our work focuses on improving the performance of GEC models in OOD scenarios. Specifically, we show that GEC models struggle in OOD settings because the errors they learn to correct during training do not generalize to new domains. NEs, due to their long-tail nature, are easily memorized by ASR models. For Whisper-like open-access models, we do not know what NEs were seen during training and whether an NE encountered during inference is unseen by the model. One of the primary motivations of DARAG is to improve on named entities never seen before (which is also challenging as they cannot be corrected with linguistic knowledge). Table 4 shows some compelling results for this, where ASR models trained on OOD datasets show significant performance boosts with DARAG-based NE retrieval. These issues are outlined in the Introduction and discussed in detail in Section 3.2.

Furthermore, as highlighted in our paper, using open-access models limits our ability to understand how GEC models operate, what they learn, and where they fail. The primary contribution of our work is to conduct controlled experiments on single datasets and OOD scenarios to identify and address the limitations of GEC methods. This approach reflects realistic industrial use cases, where ASR systems often encounter OOD data, and open-access models like Whisper are not typically employed.

F Why is SentenceBERT an effective NE Retriever?

Our choice of using SentenceBERT as a NE retriever is driven by the observation that NEs are often included in the best hypothesis generated by the ASR model but with incorrect spellings (Guo et al., 2019). This issue commonly arises due to phonetically similar or confused sounds generated by the ASR system. For example, as shown in Table 14, "Phillip" was transcribed without the "l," and "Sharon" was transcribed as "Shared." SentenceBERT excels at retrieving semantically similar words from a corpus, making it highly effective at identifying the correct NEs. It achieves a retrieval accuracy of $\geq$92% for the top k retrieved NEs.

G Rationale behind our OOD setup

Previous works on ASR OOD evaluation employ a variety of settings (Hu et al., 2024b; Seth et al., 2024). However, our primary focus is not on OOD adaptation or evaluation itself; rather, we aim to demonstrate that DARAG enhances performance in typical OOD scenarios. To this end, we have chosen to use widely recognized benchmark datasets for both ID and OOD evaluations. These datasets not only feature standard train-dev-test splits but also represent fundamentally different domains. Furthermore, while some prior works rely on synthetic datasets for their experimental setups (Hu et al., 2024b), our approach uses real-world data for evaluation, aligning more closely with the practical motivations of our study.

H Why does generating synthetic transcripts from noisy hypothesis prove to be effective?

Our hypothesis for this step is simple: We generate synthetic transcripts to capture in-domain linguistic features. These transcripts are then used to generate audios that are used to train ASR models to learn such linguistic features and domain-specific words and entities. LLMs are robust to noisy in-context exemplar (Gao et al., 2024b; Zhu et al., 2023), and the majority of the errors that arise in hypothesis only arise from spelling mistakes or mistakes in transcribing named entities. Thus, generating synthetic transcripts from LLMs proves to be a simple and robust solution for bridging the domain gap.

I In-Domain Performance in Out-of-Domain Settings

Table 9 presents the performance of DARAG on in-domain tests after augmenting the hypotheses dataset with OOD hypotheses-transcription pairs. The results demonstrate that DARAG maintains its performance on the in-domain test with only a negligible drop.

Test	Method	Train	OOD Adapt.	WER (↓)
Vox	Baseline	-	-	10.1
	+DARAG	Vox	-	8.6
	+DARAG	Vox	LS	8.9
	+DARAG	Vox	SPGI	9.0
	+DARAG	Vox	TED	9.0
LS (Other)	Baseline	-	-	8.4
	+DARAG	LS	-	6.4
	+DARAG	LS	Vox	7.5
	+DARAG	LS	SPGI	7.8
	+DARAG	LS	TED	6.9

Table 9: Performance comparison of DARAG across different settings. OOD Adapt. refers to the dataset for which synthetic data was generated and augmented to the original hypotheses for GEC training. Our results show that, even with the addition of synthetically generated training data, DARAG maintains its in-domain performance. Furthermore, improvements in a specific domain occur only when the augmentations are consistent with that domain. This approach ensures that the errors used for training match the characteristics of those the ASR model will encounter during testing.

J Hyper-parameter Tuning

J.1 Effect of k for NE retrieval

Table 10 compares the performance of DARAG across various values of k for NE retrieval. We choose two in-domain settings as our main experiments show NE retrieval is most effective in in-domain scenarios. We show both higher and lower values of k can lead to a drop in performance and find 5 as the most optimal value. Higher values of k can retrieve irrelevant NEs and confuse the GEC model. Lower values of k can lead to cases where the GT NE is not retrieved.

Test	k=1	k=2	k=5	k=7	k=9
Vox	87.8	88.7	**90.0**	87.9	87.8
LS (Other)	94.5	94.5	**96.4**	93.9	93.3

Table 10: Performance comparison of DARAG on two in-domain settings with various values of k for NE retrieval.

J.2 Effect of n_{small} in OOD settings

Table 11 compares the performance of DARAG across various values of n_{small}. Larger n_{small} can lead to more diverse and consistent augmentations, improving performance. For our primary experiments, we stick to 100 to keep our setting ultra-low-resource.

Test	10	50	100	500
Vox	15.2	11.3	10.0	**9.5**
SPGI	17.9	14.1	12.0	**11.7**

Table 11: Performance comparison of DARAG on two OOD settings (with LS as training set) with various values of n_{small}. Larger values can lead to improved performance.

J.3 Effect of n_{syn}

Table 12 compares the performance of DARAG using different values of n_{syn}, represented as a factor of n (the size of the original training set for the target dataset in an OOD setting). Increasing the number of synthetic samples (higher n_{syn}) can provide more diverse and consistent augmentations in OOD settings, resulting in better performance. However, the improvements plateau beyond a certain point. For our main experiments, we use $n_{\text{syn}} = 1$ due to resource limitations.

Test	0.5×	1×	2×	5×
Vox	13.1	10.0	**9.6**	9.7
SPGI	14.2	12.0	**11.3**	**11.3**

Table 12: Performance comparison of DARAG on two OOD settings (with LS as training set) across different scaling factors of n_{syn} relative to n. More synthetic samples can lead to improved performance, but plateaus beyond a certain point.

K Examples of Generated Transcripts

Table 13 provides examples of synthetically generated transcripts for each dataset from our evaluation setup. The transcripts are coherent and consistent with the characteristics of the domain.

L Examples of DARAG Corrections

Table 14 qualitatively compares DARAG with traditional GEC on various instances from benchmark datasets. We show that DARAG is able to accurately correct NEs which traditional GEC cannot. Additionally, DARAG shows superior performance in OOD scenarios.

M Additional Details

Compute details. For all our pre-training and fine-tuning experiments, we used four NVIDIA A6000-48GB GPUs. Each training requires 4-24 hours.

Potential Risk. As mentioned in the limitations section of the paper, DARAG might encode biases inherent to the LLM. This might lead to unsafe generations and corrections. Additionally, voice cloning systems used as part of our method can be employed to create deep fake voices.

Software and Packages details. We implement all our models in PyTorch [2] and use Parler-TTS [3] and LLaMa-2 [4]. We employ ESPnet (Watanabe et al., 2018) for training our ASR models.

Use of AI models. We used GPT-4 for rephrasing certain parts of the writing.

Datasets. Dataset details, together with statistics are provided below:

LibriSpeech [5] The LibriSpeech dataset is a large-scale corpus of approximately 1,000 hours of 16kHz English speech derived from audiobooks in the LibriVox project, with text sourced primarily from Project Gutenberg. It is split into training sets (100hr, 360hr, and 500hr) and dev/test sets categorized as dev clean(5hr), dev other(5hr), test clean(5hr), and test other(5hr) based on transcription difficulty. The dataset also includes n-gram language models and texts with 803 million tokens and 977,000 unique words, making it valuable for Automatic Speech Recognition (ASR) research.

SPGISpeech [6] SPGISpeech is a large-scale speech transcription dataset containing 5,000 hours of professionally transcribed financial audio, including company earnings calls with a variety of L1 and L2 English accents. It features approximately 50,000 speakers and offers high-quality transcripts that have been thoroughly edited for accuracy, including proper punctuation, capitalization, and denormalization of non-standard words. The audio is split into 5 to 15-second slices, formatted as single-channel, 16kHz, 16-bit WAV files, making it ideal for training advanced speech recognition models.

VoxPopuli [7] VoxPopuli is a large-scale multilingual speech corpus designed for tasks like representation learning, semi-supervised learning, and interpretation. It offers 400,000 hours of unlabeled speech in 23 languages, resulting in 8K-24K hours of data for each language, 1,800 hours of transcribed speech in 16 languages, and 17,300 hours of speech-to-speech interpretation across 15 language pairs. In transcribed speech, the filtered utterances are split into train, development and test sets with disjoint speakers and target duration ratio (18:1:1). We only use the English language split which has 543 hours of transcribed speech. Additionally, it includes 29 hours of transcribed non-native English speech for research on accented speech in ASR.

GigaSpeech [8] GigaSpeech is a large-scale English speech recognition corpus with 10,000 hours of training set of high-quality human-transcribed audio for supervised learning, 12 hours of dev set, and 40 hours of test set. It is designed for both supervised and unsupervised/semi-supervised learning tasks, covering a wide range of domains. It is particularly suited for large-scale speech recognition model training and adaptation.

TED-LIUM (v1) [9] The TED-LIUM corpus is a dataset of English-language TED talks, featuring transcriptions of talks sampled at 16kHz. It contains approximately 118 to 452 hours of transcribed speech data, with 56,803 examples in the training set, 1,469 in the test set, and 591 in the validation set. This dataset is widely used for Automatic Speech Recognition (ASR) research and model training.

All datasets used in our paper are openly available for download and free to use to academic research.

[2] https://pytorch.org/

[3] https://github.com/huggingface/parler-tts

[4] https://huggingface.co/meta-llama

[5] https://www.openslr.org/12

[6] https://datasets.kensho.com/datasets/spgispeech

[7] https://github.com/facebookresearch/voxpopuli

[8] https://github.com/SpeechColab/GigaSpeech

[9] https://www.openslr.org/7/

Dataset	Synthetic Transcripts
LibriSpeech	the duke entered the grand hall as the musicians began playing a lively gavotte
LibriSpeech	her highness attended the gala wearing the renowned emerald necklace from the royal collection
SPGI	Sarah, can we reassess the projected growth for the third quarter and adjust our targets accordingly?
SPGI	Our current expectation is to maintain a minimum margin of 40%, though market conditions may lead to some adjustments.
GigaSpeech	please navigate to the settings page to update your api key and configure the callback url.
GigaSpeech	she served as the vice chair of the european data protection board for three years before joining the united nations privacy task force.
VoxPopuli	as the smoke cleared the battered zeppelin drifted slowly back towards the enemy's encampment
VoxPopuli	yet i shall not yield to their demands but will defend my honor just as young frederick once did in times of great peril
TED	we are often overwhelmed by too many options and that can make even simple decisions difficult to navigate
TED	i must admit that my journey has had its ups and downs but in the end i found exactly what i was looking for

Table 13: Examples of generated transcripts by the DARAG methodology.

Dataset	ASR Transcription	Traditional GEC	DARAG
LibriSpeech Other	how eye wish you could get me a coffee of that pitcher phillip laura said in treating lee	how i wish you could get me a coffee of that pitcher phillip laura said in treatingly	how i wish you could get me a copy of that picture philip laura said treatingly
LibriSpeech Other (OOD on Vox)	but she fixed up on a pitcher which she said she preferred too anything she had scene in the galley	but she fixed up on a pitcher which she said she preferred too anything she had scene in the galley	but she fixed upon a picture which she said she preferred to anything she had seen in the gallery
SPGI	and we expect once the Sharon Nation Credit gets taken care of, we're in a arrange where we will be managing in flows and out flows on a normal	and we expect once the Sharon Nation Credit gets taken care of, we're in a arrange where we will be managing in flows and out flows on a normal	and we expect once the Shared National Credit gets taken care of, we're in a range where we will be managing inflows and outflows on a normal
SPGI (OOD on Vox)	obviously, the confidence level on future and growing exploration in the Golf of Mexico, in South East Asia. So	obviously, the confidence level on future and growing exploration in the Golf of Mexico, in South East Asia. So	obviously, the confidence level on future and growing exploration in the Gulf of Mexico, in Southeast Asia. So
GigaSpeech	TRULY THE EIGHT WONDER OF THE WORLD SEAN ELLIOT. THANK YOU SO MUCH.	TRULY THE EIGHT WONDER OF THE WORLD SEAN ELLIOT. THANK YOU SO MUCH.	TRULY THE EIGHTH WONDER OF THE WORLD SHAWN ELLIOTT . THANK YOU SO MUCH .
GigaSpeech (OOD on Vox)	MICROSOFT FIRED BACK WITH ITS OWN SEARCH INJUN	MICROSOFT FIRED BACK WITH ITS OWN SEARCH INJUN	MICROSOFT FIRED BACK WITH ITS OWN SEARCH ENGINE
VoxPopuli	we need mores sources we need mores pipes than one from rush ya	we need mores sources we need mores pipes than one from rush ya	we need more sources we need more pipes than one from russia
VoxPopuli (OOD on LibriSpeech)	may i in decay however that the protection of arbitration agreements should not limited the free circulation of judgments in the union	may i indicate however that the protection of arbitration agreements should not limited the free circulation of judgments in the union	may i indicate however that the protection of arbitration agreements should not limit the free circulation of judgements in the union

Table 14: Examples of incorrect ASR transcriptions and their corresponding corrections by DARAG.