# A Comparative Analysis of Generative Error Correction Strategies for Long-Form and Noisy Audio

**Yanna Torres Gonçalves[1], Vincenzo Silva Fadda[1], Ticiana L. Coelho da Silva[2], José A. Fernandes de Macedo[1]**

[1]Programa de Pós-Graduação em Ciências da Computação
Universidade Federal do Ceará - Campus Pici
60.440-900 – Fortaleza – CE – Brazil

[2]Programa de Pós-Graduação em Modelagem e Métodos Quantitativos
Universidade Federal do Ceará - Campus Pici
60.440-900 – Fortaleza – CE – Brazil

{yannatorres, vincenzofadda}@alu.ufc.br,
ticianalc@insightlab.ufc.br, jose.macedo@ufc.br

***Abstract.*** *Automatic Speech Recognition (ASR) degrades substantially under real-world acoustic conditions, motivating Generative Error Correction (GER) with Large Language Models (LLMs). This study evaluates Single-ASR and Multi-ASR GER, with and without retrieval-augmented generation (RAG) strategies, across 13,500 audio samples under controlled degradation. GER reduces overall Word Error Rate by up to 43.9%, reaching 45.6% under severe degradation, but increases error on clean audio due to over-correction. Maximal Marginal Relevance (MMR) retrieval outperforms random sampling under severe conditions, while Multi-ASR remains competitive but is more condition-sensitive. These results call for adaptive, input-aware correction strategies.*

## 1. Introduction

Automatic Speech Recognition (ASR) has significantly improved transcription quality across multiple domains and languages. However, deploying ASR systems in real-world environments remains challenging, as speech is rarely captured under ideal acoustic conditions [Polevoi et al. 2025]. Background noise, transmission artifacts, and channel distortions are pervasive in institutional settings and systematically degrade transcription quality in ways that are difficult to anticipate at training time.

Recent progress in Large Language Models (LLMs) introduces a new paradigm for post-ASR processing known as **Generative Error Correction (GER)**. In this approach, a language model receives the transcription produced by an ASR system and generates a refined version through text-to-text generation, correcting grammatical, lexical, and contextual errors while preserving the original meaning [Yang et al. 2023, Ma et al. 2025]. Unlike traditional rescoring methods, GER allows the model to reason over the full sentence context and leverage the broad linguistic knowledge acquired during pretraining. Further, combining outputs from multiple ASR systems as input to the correction model has been shown to provide complementary corrective signals, improving robustness beyond what any single system can achieve [Ma et al. 2025].

Despite these advances, the interaction between acoustic degradation and generative error correction remains largely unexplored. Most existing studies evaluate GER

under relatively clean or controlled conditions, leaving open the question of whether its benefits persist as noise intensity and complexity increase. In particular, it is unclear whether degraded input limits the corrective capacity of LLMs or whether generative approaches can compensate for severe transcription errors. To our knowledge, this is the first study to systematically evaluate generative error correction under controlled combinations of acoustic degradation types and severity levels.

This work addresses this gap by systematically evaluating GER strategies under controlled acoustic degradation. Building on a dataset of 900 recordings in Brazilian Portuguese, we construct an augmented evaluation set of 13,500 audio files by applying combinations of three degradation types at two severity levels, simulating telephone transmission, bitcrushing artifacts, and background noise. We evaluate three state-of-the-art ASR systems individually and in combination, assessing both Single-ASR and Multi-ASR correction pipelines with and without Retrieval-Augmented Generation (RAG).

Our study is guided by the following research questions:

- **RQ1:** To what extent does generative error correction improve the transcription accuracy of single ASR systems?
- **RQ2:** Does combining the outputs of multiple ASR systems as input to a generative error correction model yield better transcription accuracy than correcting a single ASR output?
- **RQ3:** Does retrieval-augmented generation improve the performance of generative error correction?
- **RQ4:** How does acoustic noise degradation affect the transcription accuracy of ASR systems and generative error correction models?

The remainder of this paper is organized as follows. Section 2 reviews related work on ASR error correction and LLM-based post-processing approaches. Section 3 describes the dataset, augmentation pipeline, ASR models, and correction strategies. Section 4 presents and discusses the experimental results. Finally, Section 5 summarizes the findings and outlines directions for future work.

## 2. Related Work

Recent advances in Large Language Models (LLMs) have renewed interest in post-ASR text modeling. Unlike traditional language model rescoring, which operates over fixed ASR hypotheses, LLM-based approaches formulate error correction as a generative text-to-text task, enabling richer contextual reasoning and semantic refinement [Yang et al. 2023].

Several studies have demonstrated the effectiveness of LLMs for ASR error correction. [Yang et al. 2024] introduced the Generative Speech Error Correction (GenSEC) challenge, showing that LLMs can outperform conventional rescoring methods by leveraging in-context learning and broader linguistic knowledge. Similarly, [Yang et al. 2023] proposed task-activating prompting strategies, demonstrating that frozen LLMs can perform effective second-pass correction without task-specific fine-tuning.

More recent work has explored the use of multiple ASR hypotheses and system outputs. [Ma et al. 2025] investigated both supervised and zero-shot correction settings,

highlighting the importance of leveraging N-best lists and proposing constrained decoding strategies to control generation behavior. Their findings also suggest that combining outputs from multiple ASR systems provides complementary information, improving correction robustness through implicit ensembling.

Beyond English, [Wei et al. 2025] introduced the ASR-EC benchmark for Chinese ASR error correction and evaluated prompting, fine-tuning, and multimodal approaches. Their results indicate that prompting alone may be insufficient for robust correction, while model adaptation and multimodal signals lead to more consistent improvements.

Despite these advances, most prior work evaluates generative error correction under relatively clean or controlled acoustic conditions. The impact of real-world acoustic degradation (e.g., background noise, channel distortion, and compression artifacts) on LLM-based correction remains largely unexplored. While the effects of such degradations on ASR performance are well documented, their interaction with generative post-processing has not been systematically studied.

This work addresses this gap by providing a controlled evaluation of generative error correction strategies across multiple types and levels of acoustic degradation, further incorporating retrieval-augmented generation as a mechanism to improve robustness under noisy conditions.

## 3. Methodology

This section describes the dataset, augmentation pipeline, ASR systems, and generative error correction strategies used to address the research questions. Figure 1 illustrates the overall pipeline, which consists of four stages: (i) data preparation, (ii) ASR transcription, (iii) post-processing via generative error correction, and (iv) evaluation using Word Error Rate (WER). Each stage is detailed in the following subsections.

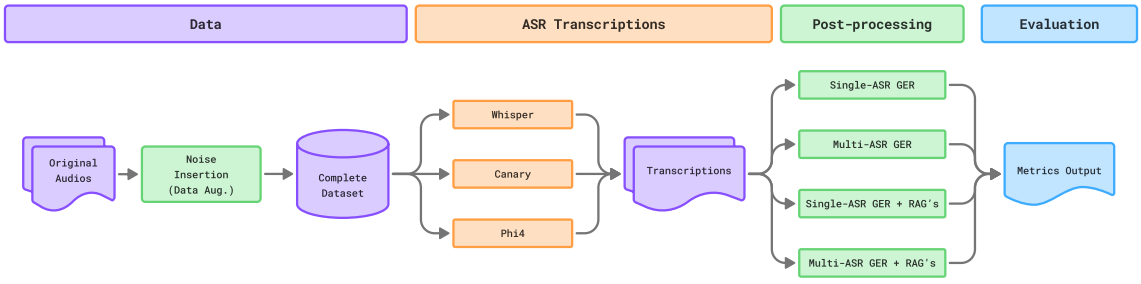



**Figure 1. Methodology Overview**

### 3.1. Dataset Construction and Augmentation

The dataset used in this study consists of 900 audio recordings in Brazilian Portuguese of medical anamnesis, sourced from a private database of Hapvida NotreDame Intermédica. The recordings were produced by three medical students (two female and one male). Audio duration ranges from 3.8 to 350.4 seconds, with a mean duration of 80.5 seconds (SD = 45.9) and a median of 74.1 seconds, totaling approximately 20.1 hours of speech content.

To evaluate the robustness of ASR systems and error correction strategies under realistic acoustic degradation conditions, a data augmentation pipeline was developed[1] to simulate common imperfections found in real-world recordings, particularly in clinical environments. Three categories of acoustic degradation were defined, each identified by a shorthand used throughout this study:

- **Telephone simulation (*tel*)**: Bandpass filtering and downsampling simulate a low-quality phone call. Level 1 (moderate): 6 kHz, 300–2400 Hz band. Level 2 (severe): 4 kHz, 300–1500 Hz band.
- **Bitcrushing (*bit*)**: Reduced bit depth and sampling rate simulate compression artifacts. Level 1 (moderate): 2 kHz with coarse quantization. Level 2 (severe): 1 kHz with heavier quantization, producing a robotic, heavily distorted output.
- **Background noise (*bgn*)**: Real-world noise from MS-SNSD [Reddy et al. 2019] is overlaid at a target Signal-to-Noise Ratio (SNR). Level 1 (moderate): 5–15 dB, voice remains intelligible. Level 2 (severe): −5–5 dB, noise and voice reach comparable volumes.

All possible combinations of the three degradation types were considered across both severity levels, yielding 14 distinct augmentation conditions per audio. Combined conditions are expressed as concatenations of the respective identifiers (e.g., *tel*+*bgn*). Applied independently to the 900 original recordings, this procedure produced 12,600 augmented samples, resulting in a total dataset of **13,500 audio files** and approximately 301 hours of speech content.

### 3.2. Automatic Speech Recognition Models

Three state-of-the-art ASR models were selected for evaluation: OpenAI Whisper [Radford et al. 2022], NVIDIA Canary [Sekoyan et al. 2025], and Microsoft Phi-4 [Abouelenin et al. 2025]. The selection was guided by their ranking on the Hugging Face Open ASR Leaderboard for Portuguese [Srivastav et al. 2025] at the time of this study, which benchmarks open-source multilingual ASR systems on standardized datasets. Each model was applied independently to the full set of audio recordings, producing baseline transcriptions used both for standalone evaluation and as input to the generative error-correction stage.

**Whisper.** Follows a Transformer encoder-decoder architecture trained on 680,000 hours of weakly supervised multilingual speech data [Radford et al. 2022]. The model addresses recognition and translation within a unified sequence-to-sequence framework, exhibiting strong zero-shot generalization across languages and domains.

**Canary.** Combines a FastConformer encoder with a Transformer decoder, supporting both speech recognition and translation across 25 languages [Sekoyan et al. 2025]. The model was trained on approximately 1.7 million hours of multilingual speech and is designed to achieve high accuracy while remaining computationally efficient.

**Phi-4.** Phi-4-Multimodal extends the Phi-4-Mini language model with a dedicated speech encoder connected through modality-specific LoRA adapters [Abouelenin et al. 2025]. This design enables speech-to-text generation within a unified multimodal architecture while preserving the underlying language model capabilities.

[1] `https://github.com/yanna-torres/audio-noise-augmentation`

### 3.3. Generative Error Correction Strategies

Two Generative Error Correction (GER) strategies are evaluated in this study, differing in the number of ASR hypotheses provided to the model as input. In **Single-ASR GER**, the LLM receives a single transcription from one ASR system and is instructed to correct only grammatical and spelling errors, without rewriting or reinterpreting the content, allowing the model's performance to be evaluated in isolation. In **Multi-ASR GER**, the LLM receives one transcription from each of the three ASR systems, in randomized order and without indication of source, and is instructed to compare them, identify the most accurate parts of each, and merge them into a single unified transcription, exploiting complementary capabilities while minimizing system-specific bias.

All generative error correction experiments were performed using Qwen3-4B-Instruct [Yang et al. 2025]. The model was prompted to act as a professional medical transcription editor, with explicit instructions to preserve the original wording, maintain correct Brazilian Portuguese grammar and spelling, and avoid adding or removing information beyond what was present in the input transcriptions. Hallucination artifacts in the form of looping repetitions, a known failure mode of ASR systems, were also addressed in the prompt by instructing the model to consider only the first occurrence of repeated segments. In both settings, the model was required to briefly explain its reasoning before producing the final output, which was delimited by a structured marker to facilitate automated extraction; the Multi-ASR prompt additionally instructed the model to cross-reference the three input transcriptions and justify segment-level choices.

### 3.4. Retrieval-Augmented Generation (RAG)

To further enhance both GER strategies, a Retrieval-Augmented Generation (RAG) mechanism was incorporated to provide contextual examples to the generative model during correction. These examples consist of previously curated medical transcription texts representative of well-formed clinical narratives.

The retrieval pool comprised 2,700 reference texts, randomly sampled from the same source database as the audio recordings, without filtering by audio duration, anamnesis type, or other content criteria. These texts were kept strictly separate from the audio samples used in evaluation, ensuring that no evaluation instance appears in the retrieval set. This separation prevents data leakage while still exposing the model to representative language, terminology, and narrative structures characteristic of the target domain.

For each correction prompt, $K = 5$ examples were retrieved from this pool and included as contextual demonstrations alongside the candidate transcription. The retrieved examples were prepended to the user prompt as contextual demonstrations preceding the transcription input, explicitly instructing the model to use them as reference for terminology consistency and correction patterns. Two retrieval strategies were evaluated across both settings, as described below.

**Random Sampling.** In this strategy, $K$ examples are drawn uniformly at random from the retrieval pool for each correction prompt. This approach makes no assumption about the relevance or diversity of the selected examples and serves as a simple, unbiased baseline against which more sophisticated retrieval methods can be compared.

**Maximal Marginal Relevance (MMR).** MMR selects examples by iteratively

balancing two competing criteria: relevance to the input and diversity among the already selected examples [Goldstein and Carbonell 1998]. Both the query (the candidate transcription) and the 2,700 pool texts were embedded using Qwen3-Embedding-4B [Zhang et al. 2025], with embeddings normalized and compared via cosine similarity. Examples were selected iteratively: the first example is the one with highest similarity to the query, and each subsequent example $d$ is chosen from the remaining candidates to maximize

$$\mathrm{MMR}(d) = \lambda \cdot \mathrm{sim}(d, q) - (1 - \lambda) \cdot \max_{d_j \in S} \mathrm{sim}(d, d_j), \tag{1}$$

where $q$ is the query, $S$ is the set of already-selected examples, and $\lambda \in [0, 1]$ controls the trade-off between relevance and diversity. We set $\lambda = 0.7$, favoring relevance while still penalizing redundancy among the retrieved examples. This mechanism reduces redundancy in the retrieved context while maintaining topical relevance, producing a more informative and varied set of demonstrations for the generative model.

### 3.5. Evaluation Metric

Word Error Rate (WER) was adopted as the evaluation metric for transcription quality. WER is the standard metric for ASR systems, quantifying the proportion of incorrectly transcribed words relative to the total number of words in the reference. Errors are categorised as insertions ($I$), substitutions ($R$), and deletions ($D$), with correct words counted as hits ($H$). Formally, WER is defined as:

$$WER = \frac{I + R + D}{H + R + D} \tag{2}$$

WER values may exceed 1.0 under severe degradation conditions, where the number of errors surpasses the number of reference words. The lower the WER, the better the transcription.

## 4. Experimental Results

This section presents the results of the 15 experiments evaluated across 13,500 audio files, organized according to the research questions defined in Section 1.

As shown in Figure 2, the evaluated strategies exhibit distinct performance profiles across configurations, with clear differences between baseline ASR, generative correction, and retrieval-based approaches.

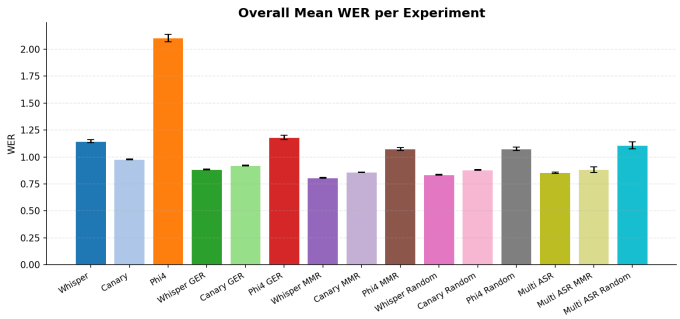



**Figure 2. Overall WER across all evaluated experiments**

### 4.1. Baseline ASR Performance

Table 1 presents the overall mean WER for all 15 experiments across the full evaluation set, including original and degraded conditions. Among the three baseline ASR systems, Canary achieved the lowest overall WER (0.978), followed by Whisper (1.146) and Phi4 (2.105). On the original, non-degraded recordings, the ranking changes notably: Whisper achieves the lowest WER (0.216), followed by Canary (0.270) and Phi4 (0.430). This indicates that Whisper's strong zero-shot performance on clean speech does not translate proportionally to noisy conditions, while Canary demonstrates more consistent behaviour across acoustic conditions.

**Table 1. WER results per experiment across degradation conditions. Bold indicates the best result per column; underline indicates the second best.**

| Experiment | Original | 1 deg. | 2 deg. | 3 deg. | Overall |
|---|---|---|---|---|---|
| Whisper | **0.216** | 1.006 | 1.414 | 1.226 | 1.146 |
| Canary | 0.270 | 0.847 | 1.154 | 1.200 | 0.978 |
| Phi4 | 0.430 | 1.851 | 2.524 | 2.450 | 2.105 |
| Whisper GER | 0.399 | 0.823 | 0.985 | **0.979** | 0.883 |
| Canary GER | 0.463 | 0.840 | 1.028 | 1.079 | 0.922 |
| Phi4 GER | 0.524 | 1.076 | 1.352 | 1.318 | 1.182 |
| Whisper MMR | 0.313 | **0.722** | **0.917** | <u>0.983</u> | **0.807** |
| Whisper Random | 0.358 | 0.755 | <u>0.941</u> | 0.990 | <u>0.834</u> |
| Canary MMR | 0.347 | 0.769 | 0.975 | 1.037 | 0.859 |
| Canary Random | 0.431 | 0.802 | 0.983 | 1.025 | 0.880 |
| Phi4 MMR | 0.448 | 0.991 | 1.220 | 1.202 | 1.074 |
| Phi4 Random | 0.488 | 0.993 | 1.215 | 1.200 | 1.076 |
| Multi ASR | 0.386 | 0.782 | 0.961 | 0.990 | 0.855 |
| Multi ASR MMR | <u>0.265</u> | <u>0.743</u> | 1.042 | 1.139 | 0.883 |
| Multi ASR Random | 0.304 | 0.865 | 1.317 | 1.606 | 1.107 |

### 4.2. Effect of Generative Error Correction on Single-ASR Systems (RQ1, RQ3)

Comparing each ASR system against its GER variant reveals a nuanced pattern. On the overall evaluation set, GER consistently reduces WER for Whisper (1.146 $\rightarrow$ 0.883) and Phi4 (2.105 $\rightarrow$ 1.182), representing substantial improvements of 22.9% and 43.9%, respectively. For Canary, the reduction is more modest (0.978 $\rightarrow$ 0.922, 5.8%). However, on original audio only, GER consistently *increases* WER across all three systems: Whisper (0.216 $\rightarrow$ 0.399), Canary (0.270 $\rightarrow$ 0.463), and Phi4 (0.430 $\rightarrow$ 0.524). This pattern indicates that GER provides the greatest benefit under degraded conditions, where transcription errors are more frequent and the model has more opportunities for correction. On clean speech, the correction process introduces unnecessary modifications that increase error rates, suggesting a tendency toward over-correction when the input transcription is already of high quality.

Augmenting GER with retrieval-augmented generation further reduces WER across all three ASR systems. For Whisper, MMR retrieval achieves the lowest overall WER among all single-ASR configurations (0.807), outperforming both GER alone

(0.883) and random retrieval (0.834). A similar pattern holds for Canary (MMR: 0.859, Random: 0.880, GER: 0.922) and Phi4 (MMR: 1.074, Random: 1.076, GER: 1.182). The consistent advantage of MMR over random sampling across all three systems suggests that retrieval diversity, as promoted by the MMR strategy, provides more informative contextual guidance than unstructured random selection. On original audio, however, RAG does not resolve the over-correction problem observed in plain GER, with WER remaining elevated relative to the ASR baselines across all configurations.

### 4.3. Multi-ASR Generative Error Correction (RQ2, RQ3)

The Multi-ASR baseline, which provides all three ASR outputs to the correction model without retrieval, achieves an overall WER of 0.855, remaining competitive with the best single-ASR RAG configurations and outperforming all plain GER variants. On original audio, Multi-ASR MMR achieves the lowest WER among all corrected configurations (0.265), although it does not surpass the best raw ASR result (Whisper: 0.216). Notably, Multi-ASR Random performs considerably worse than both Multi-ASR and Multi-ASR MMR in the overall evaluation (1.107 vs. 0.855 and 0.883), suggesting that unstructured retrieval is particularly detrimental in the multi-ASR setting, where the model already receives richer and more complex input. These results indicate that combining multiple ASR outputs yields competitive performance, but its effectiveness is highly sensitive to the retrieval strategy employed.

A closer inspection of these results suggests that the effectiveness of Multi-ASR correction depends on the balance between complementary and conflicting information across system outputs. When individual ASR systems produce partially correct but diverse transcriptions, the model can exploit this variability to recover missing or erroneous segments. However, under severe degradation, multiple systems often produce similarly corrupted or inconsistent outputs, thereby reducing the benefit of aggregation and increasing ambiguity. In such cases, retrieval-based guidance becomes more critical, as it provides external grounding that helps stabilize the correction process. This interaction indicates that Multi-ASR correction is inherently condition-dependent, with its effectiveness shaped by both input diversity and retrieval quality.

### 4.4. Impact of Acoustic Degradation (RQ4)

As illustrated in Figure 3, WER increases consistently with the number of simultaneously applied degradations across all 15 experiments, confirming a systematic relationship between degradation complexity and transcription error. For Whisper, WER rises from 0.216 on original audio to 1.006 under single degradation and 1.414 under two simultaneous degradations. Phi4 exhibits the steepest degradation curve, with WER reaching 2.524 under two simultaneous degradations, reflecting its higher baseline sensitivity to noise. GER and RAG variants consistently attenuate but do not eliminate degradation effects, with the gap between corrected and uncorrected systems narrowing as degradation severity increases.

The three degradation types produce markedly different impacts on transcription accuracy. Telephone simulation (*tel*) is the least damaging condition across all systems, with Whisper and Canary achieving WERs of 0.480 and 0.523, respectively, under single telephone degradation, values that remain close to their clean baselines. This is consistent with the nature of the degradation: bandpass filtering preserves the mid-frequency range

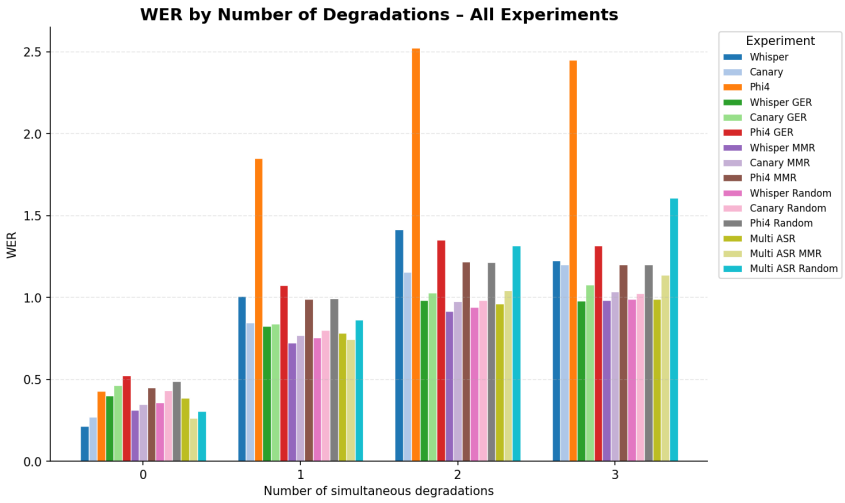
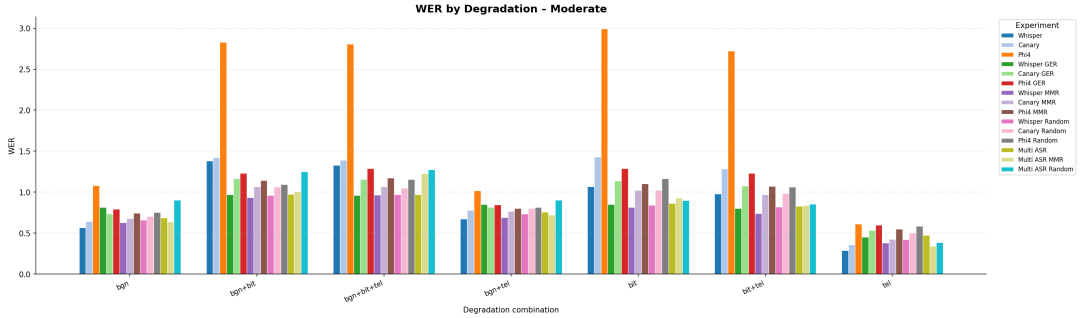



**Figure 3. WER vs. number of applied degradations for all experiments**

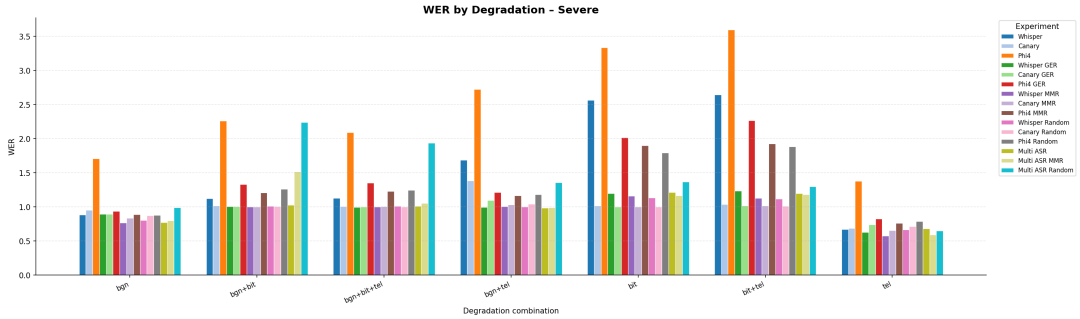
most relevant to speech intelligibility, leaving a sufficiently informative acoustic signal for transcription. Bitcrushing (*bit*), by contrast, is the most destructive single degradation, with Whisper reaching 1.814 and Phi4 an extreme 3.164. Background noise (*bgn*) occupies an intermediate position, with Whisper reaching 0.726 and Canary 0.796 under single background noise degradation, where the voice signal remains present but with reduced prominence relative to the noise floor. Combined degradations amplify these effects in a non-linear manner, with *bit+tel* and *bit+bgn* producing some of the highest WERs in the evaluation. Severity level also has a pronounced and asymmetric effect: for bitcrushing, Whisper WER rises from 1.066 at Level 1 to 2.563 at Level 2, while for telephone simulation the increase is more gradual, from 0.290 to 0.671, underscoring that not all degradations scale equally with severity.

In addition to degradation type, severity level introduces a distinct and consistent effect on transcription accuracy. Across all systems, as shown in Figure 4, increasing degradation from Level 1 to Level 2 leads to a substantial rise in WER, although the magnitude of this increase varies by degradation type. Bitcrushing exhibits the most pronounced sensitivity to severity, with errors increasing sharply between levels, while telephone simulation produces a more gradual degradation. Background noise shows intermediate behaviour, where increased noise levels reduce intelligibility without completely disrupting the signal structure. These patterns indicate that severity is not merely a scaling factor, but interacts with the nature of the degradation, amplifying different failure modes depending on the underlying distortion.

GER strategies show consistent but degradation-dependent benefits. Under telephone simulation, where baseline WER is already modest, all correction strategies provide limited absolute gains. Under bitcrushing, GER yields its largest improvements: Whisper GER reduces *bit* WER from 1.814 to 1.041, and Whisper MMR further to 0.987, representing a relative reduction of 45.6% over the baseline. A similar pattern holds for Phi4, where GER reduces *bit* WER from 3.164 to 1.651 and MMR to 1.502. Under background noise, gains are moderate and consistent across all correction strategies. Across all degradation types, MMR retrieval outperforms random sampling, and its advantage is most pronounced under severe conditions, suggesting that retrieval diversity becomes




**(a) Level 1**




**(b) Level 2**

**Figure 4. WER by experiment across degradation types for two severity levels.**

more valuable as the input signal degrades.

The optimal correction strategy varies by degradation type, as shown in Table 2. Under telephone simulation, where individual transcriptions remain relatively reliable, Multi-ASR MMR achieves the lowest *tel* WER (0.465), indicating that cross-system diversity is more valuable than retrieval alone when the acoustic signal is largely preserved. Under bitcrushing, single-ASR MMR consistently outperforms Multi-ASR variants (Whisper MMR achieves 0.987 versus Multi ASR's 1.039) likely because the degraded outputs produced by multiple systems introduce conflicting signals to the correction model, whereas a single curated retrieval context provides more stable grounding. Under background noise, Whisper MMR achieves the best result among single-ASR configurations (0.698), while Multi ASR MMR (0.718) remains competitive. These patterns suggest that no single strategy dominates across all conditions: retrieval-augmented single-ASR correction is preferable under severe degradation, while multi-system combination with MMR retrieval is more effective when degradation is moderate and individual transcriptions retain partial intelligibility.

Beyond aggregate WER, qualitative analysis of the transcriptions reveals distinct failure modes for each ASR system under severe degradation. Whisper exhibits hallucination behaviour under bitcrushing conditions, generating looping content unrelated to the audio input, including repeated phrases such as *"Cidade do Brasil"*, *"Thank you. Thank you for watching!"*, and sequences of repeating numbers. This behaviour is consistent with previously reported hallucination tendencies in Whisper under low-signal conditions [Radford et al. 2022], where the model defaults to generating plausible but ungrounded content when the acoustic signal is too degraded. Canary, by contrast, fails silently under

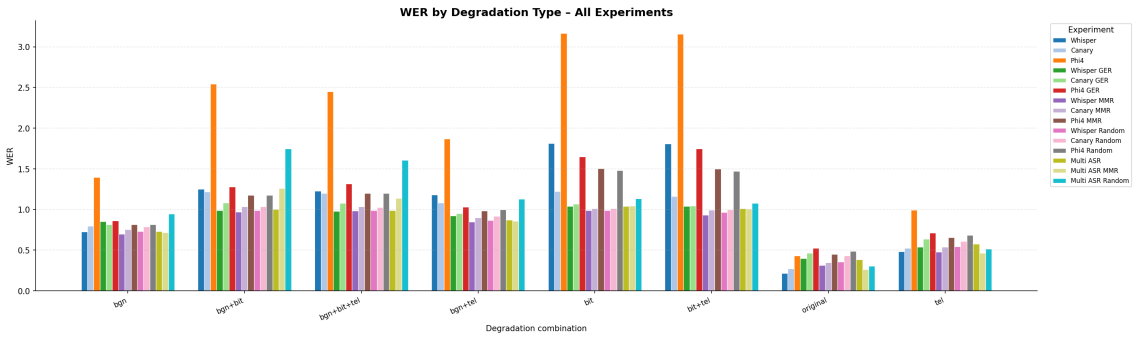



**Figure 5. WER across degradation types for all experiments**

**Table 2. Mean WER per experiment by degradation type. Bold indicates the best result per column; underline indicates the second best.**

| Experiment | tel | bit | bgn | tel+bit | tel+bgn | bit+bgn | tel+bit+bgn |
|---|---|---|---|---|---|---|---|
| Whisper | <u>0.480</u> | 1.814 | 0.726 | 1.810 | 1.178 | 1.253 | 1.226 |
| Canary | 0.523 | 1.223 | 0.796 | 1.161 | 1.082 | 1.219 | 1.200 |
| Phi4 | 0.994 | 3.164 | 1.394 | 3.158 | 1.870 | 2.545 | 2.450 |
| Whisper GER | 0.540 | 1.041 | 0.854 | 1.039 | 0.923 | 0.988 | **0.979** |
| Canary GER | 0.638 | 1.069 | 0.814 | 1.046 | 0.953 | 1.084 | 1.079 |
| Phi4 GER | 0.713 | 1.651 | 0.864 | 1.749 | 1.029 | 1.280 | 1.318 |
| Whisper MMR | <u>0.480</u> | **0.987** | **0.698** | **0.934** | **0.848** | **0.968** | <u>0.983</u> |
| Whisper Random | 0.544 | <u>0.989</u> | 0.732 | <u>0.967</u> | 0.868 | <u>0.987</u> | 0.990 |
| Canary MMR | 0.540 | 1.011 | 0.755 | 0.992 | 0.898 | 1.034 | 1.037 |
| Canary Random | 0.607 | 1.013 | 0.787 | 0.996 | 0.920 | 1.033 | 1.025 |
| Phi4 MMR | 0.655 | 1.502 | 0.817 | 1.499 | 0.985 | 1.176 | 1.202 |
| Phi4 Random | 0.686 | 1.478 | 0.815 | 1.473 | 0.998 | 1.176 | 1.200 |
| Multi ASR | 0.578 | 1.039 | 0.730 | 1.013 | 0.869 | 1.000 | 0.990 |
| Multi ASR MMR | **0.465** | 1.046 | <u>0.718</u> | 1.008 | <u>0.856</u> | 1.261 | 1.139 |
| Multi ASR Random | 0.518 | 1.132 | 0.945 | 1.076 | 1.130 | 1.745 | 1.606 |

bitcrushing, returning empty transcriptions when the audio signal falls below a minimum intelligibility threshold. Phi4 exhibits a different failure mode, producing the minimal output *"A."* for audio files it cannot process, suggesting an internal fallback mechanism rather than a gradual degradation of transcription quality. These qualitative differences highlight that aggregate WER alone does not fully characterise ASR robustness under severe noise, and that the nature of system failures is as informative as their frequency.

## 5. Conclusion and Future Works

This work presented a systematic evaluation of generative error correction strategies for long-form medical audio under controlled acoustic degradation. By analyzing 15 experimental configurations across 13,500 audio samples, we investigated how correction strategies interact with varying levels and types of noise.

The results demonstrate that generative error correction is highly effective under degraded conditions, significantly reducing transcription errors when the input signal is

corrupted. However, this benefit does not extend to clean audio, where correction models tend to introduce unnecessary modifications, leading to over-correction. This highlights a fundamental limitation of generative approaches: their effectiveness depends strongly on the initial quality of the transcription.

The incorporation of retrieval-augmented generation further improves correction performance, particularly when using diversity-aware strategies such as maximal marginal relevance. These methods provide more informative contextual grounding than random sampling, especially under severe degradation, where input transcriptions are less reliable. In contrast, unstructured retrieval can negatively impact performance, particularly in Multi-ASR settings.

Combining multiple ASR outputs provides competitive performance, but its effectiveness is condition-dependent. While cross-system diversity can improve correction when transcriptions contain complementary information, it becomes less beneficial under severe degradation, where multiple systems tend to produce similarly corrupted outputs. In such scenarios, retrieval-based grounding plays a more critical role in stabilizing the correction process.

Finally, the analysis of acoustic degradation reveals that transcription performance is shaped not only by the presence of noise, but also by its type and severity. Bitcrushing emerges as the most disruptive condition, while telephone simulation has a comparatively limited impact. Severity levels further modulate these effects, indicating that degradation is not a uniform factor but interacts with the underlying distortion mechanisms.

These findings suggest that effective post-ASR correction requires adaptive strategies that account for both acoustic conditions and input quality, rather than relying on a single correction pipeline.

As future work, we plan to explore adaptive correction strategies that dynamically select between Single-ASR and Multi-ASR inputs based on estimated transcription quality. Additionally, we aim to investigate fine-tuned or domain-adapted language models to reduce over-correction on clean audio, as well as more robust retrieval mechanisms that better align contextual examples with degraded inputs.

**Acknowledge.** This research was conducted as part of the CEREIA project and was supported by multiple institutions and partners. We gratefully acknowledge the contributions from Hapvida NotreDame Intermédica, the Federal University of Ceará (UFC), and the São Paulo Research Foundation (FAPESP). This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number #2025/12065-7. We extend our gratitude to all the institutions, partners, and funders involved in this project.

**Generative AI Use Disclosure** Generative AI was used as an editing and language refinement tool to improve clarity, grammar, and organization of the manuscript. It was not used to generate experimental results, data analyses, or core scientific contributions. All methodological design, implementation, experiments, and interpretations were conducted by the authors.

## References

Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., et al. (2025). Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Goldstein, J. and Carbonell, J. G. (1998). Summarization:(1) using mmr for diversity-based reranking and (2) evaluating summaries. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 181–195.

Ma, R., Qian, M., Gales, M., and Knill, K. (2025). Asr error correction using large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1389–1401.

Polevoi, A., Kragin, A., and Loukachevitch, N. (2025). Ground truth-free wer prediction for asr via audio quality and model confidence features. In *International Conference on Speech and Computer*, pages 29–44. Springer.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.

Reddy, C. K., Beyrami, E., Pool, J., Cutler, R., Srinivasan, S., and Gehrke, J. (2019). A scalable noisy speech dataset and online subjective test framework. In *Interspeech 2019*, pages 1816–1820. ISCA.

Sekoyan, M., Koluguri, N. R., Tadevosyan, N., Zelasko, P., Bartley, T., Karpov, N., Balam, J., and Ginsburg, B. (2025). Canary-1b-v2 & parakeet-tdt-0.6 b-v3: Efficient and high-performance models for multilingual asr and ast. *arXiv preprint arXiv:2509.14128*.

Srivastav, V., Zheng, S., Bezzam, E., Bihan, E. L., Koluguri, N., Żelasko, P., Majumdar, S., Moumen, A., and Gandhi, S. (2025). Open asr leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation.

Wei, V. J., Wang, W., Jiang, D., Song, Y., and Wang, L. (2025). Asr-ec benchmark: Evaluating large language models on chinese asr error correction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1567–1575.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. (2025). Qwen3 technical report.

Yang, C.-H. H., Gu, Y., Liu, Y.-C., Ghosh, S., Bulyko, I., and Stolcke, A. (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.

Yang, C.-H. H., Park, T., Gong, Y., Li, Y., Chen, Z., Lin, Y.-T., Chen, C., Hu, Y., Dhawan, K., Żelasko, P., et al. (2024). Large language model based generative error correction: A challenge and baselines for speech recognition, speaker tagging, and emotion recognition. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 371–378. IEEE.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., and Zhou, J. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.