MEDSAGE: Enhancing Robustness of Medical Dialogue Summarization to ASR Errors with LLM-generated Synthetic Dialogues

Kuluhan Binici[1,2,3*], **Abhinav Ramesh Kashyap**[4†], **Viktor Schlegel**[5,6†], **Andy T. Liu**[1†], **Vijay Prakash Dwivedi**[7†], **Thanh-Tung Nguyen**[1†], **Xiaoxue Gao**[8], **Nancy F. Chen**[8], **Stefan Winkler**[1,3]

[1]ASUS Intelligent Cloud Services (AICS)
[2]SAP
[3]National University of Singapore
[4]Crayon Software
[5]Imperial College London, Imperial Global Singapore
[6]University of Manchester, Department of Computer Science
[7]Stanford University
[8]Institute for Infocomm Research, A*STAR

Abstract

Automatic Speech Recognition (ASR) systems are used to transcribe speech into text, yet the errors they introduce can significantly degrade the performance of downstream tasks like summarization. This issue is particularly pronounced in clinical dialogue summarization, a low-resource domain where supervised data for fine-tuning is scarce, necessitating the use of ASR models as black-box solutions. Employing conventional data augmentation for enhancing the noise robustness of summarization models is not feasible either due to the unavailability of sufficient medical dialogue audio recordings and corresponding ASR transcripts. To address this challenge, we propose MEDSAGE, an approach for generating synthetic samples for data augmentation using Large Language Models (LLMs). Specifically, we leverage the in-context learning capabilities of LLMs and instruct them to generate ASR-like errors based on a few available medical dialogue examples with audio recordings. Experimental results show that LLMs can effectively model ASR noise, and incorporating this noisy data into the training process significantly improves the robustness and accuracy of medical dialogue summarization systems. This approach addresses the challenges of noisy ASR outputs in critical applications, offering a robust solution to enhance the reliability of clinical dialogue summarization.

1 Introduction

Automatic Speech Recognition (ASR) (Yu and Deng 2016) is the task of transcribing speech signals into text, enabling a wide range of applications from voice-activated assistants to automated customer service systems. ASR systems significantly aid various downstream tasks such as dialogue summarization (Liu et al. 2019; Zhong et al. 2022), where the goal is to distill key information from spoken interactions. However, errors introduced by ASR systems can degrade the performance of these summarization tasks (Li et al. 2014; Guo et al. 2024), which limits their application in high-stake domains where correctness of the summaries is important. The synthesis of Electronic Medical Records (EMRs) from doctor-patient dialogues (Krishna et al. 2021) is among such summarization tasks where accuracy is critical. Here, ASR errors can lead to inaccuracies in transcribing clinical terms, medication, or procedure names, resulting in erroneous medical notes (Hodgson and Coiera 2016), which can lead to misdiagnosis, incorrect treatment plans, and potentially harmful patient outcomes.

One way to improve dialogue summarization is to improve the ASR systems. However, this requires large amounts of supervised data (Radford et al. 2022), which is often unavailable in the healthcare domain due to privacy and ethical concerns. Consequently, clinical summarization systems often treat ASR systems as black boxes, mandating that overall improvements must arise either from post-processing ASR outputs, or making down-stream methods robust to ASR errors. Recent works propose post-processing ASR outputs using LLMs to correct erroneous ASR transcripts (Radhakrishnan et al. 2023; Bai et al. 2024). Nonetheless, based on prior studies, using prompting techniques to clear noise only proves effective only when using LLMs of large sizes, typically those that exceed 100B parameters (Yang et al. 2023). Moreover, LLMs are prone to hallucinations (Maynez et al. 2020; Nagar et al. 2024), which can introduce irrelevant symptoms or medication names into the dialogue transcript, potentially degrading summarization quality while attempting to clear noise.

Alternatively, the summarization robustness to ASR errors can be improved through data augmentation techniques (Fabbri et al. 2021). However, the conventional augmentation approach of exposing the summarization model to erroneous ASR dialogues during the training phase is not feasible either, again due to the limited availability of medical dialogue audio preventing the generation of ASR dialogue transcripts (Nanayakkara et al. 2022). Heuristic approaches for augmentation, such as randomly applying a set of corruption operations, are not ideal either, as they fail to ac-

[*]Author was an intern at AICS.
[†]Authors were employees of AICS.

curately mimic ASR errors both qualitatively and quantitatively, causing the augmented training data distribution to diverge from the real test distribution (Wang et al. 2020).

Recognizing these limitations, we propose the use of LLMs to generate synthetic dialogues mimicking *real* ASR transcriptions with their characteristic errors, as a means for data augmentation. To circumvent hallucinations, we rely on the signal from downstream tasks during fine-tuning on these generated transcripts. As such, if new domain-specific entities are mistakenly hallucinated during the augmentation process, the fine-tuning phase ensures that the model learns to disregard these irrelevant entities, increasing the feasibility of this approach as a result. To accommodate for the scarcity of medical audio recordings, we leverage the in-context learning (Brown et al. 2020) capabilities to specialize LLMs for the task of ASR transcript generation based on a few descriptive examples. Specifically, using Primock57 dataset (Papadopoulos Korfiatis et al. 2022), which includes audio recordings of clinical visits alongside their corresponding human-transcribed text, we first produce noisy transcriptions using ASR models. These noisy ASR dialogue transcripts are then paired with their clean human-transcribed versions to form the few-shot examples needed for effective in-context learning.

While in-context learning enables the qualitative approximation of ASR errors such as phonetic confusions, ensuring that synthetic errors are quantitatively similar remains a challenge (Everson et al. 2024). To address this, we first analyze the error profile of ASR models by measuring their word-error-rates and the distribution of error types, including insertions, deletions, and substitutions. Subsequently, we introduce a novel description syntax that instructs the LLMs on where to make realistic errors and what types of errors to introduce. By tagging the inputs with appropriate error tags based on the measured noise profiles of the target ASR models, we ensure that the synthetic errors generated are both qualitatively and quantitatively similar to real-world ASR errors. This methodology allows us to create synthetic noisy dialogues that accurately reflect ASR error patterns, which can be used for effective data augmentation in training robust summarization models.

Our experimental evaluation reveals that large language models (LLMs) are effective noise modelers, capable of producing similar errors to those produced by ASR. This is evident from the qunatitative and qualitative similarities between the word error profiles of the synthetic dialogues we generate and actual ASR dialogue transcripts. Such similarity also reflects on the downstream summarization performance, mirroring the performance drop caused by ASR errors. Lastly, when we utilize these synthetic noisy dialogues to augment the training set of the summarization models, the performance on the noisy test set improves by up to 16%, indicating enhanced robustness against ASR errors. In summary, our contributions leading to the robustness are:

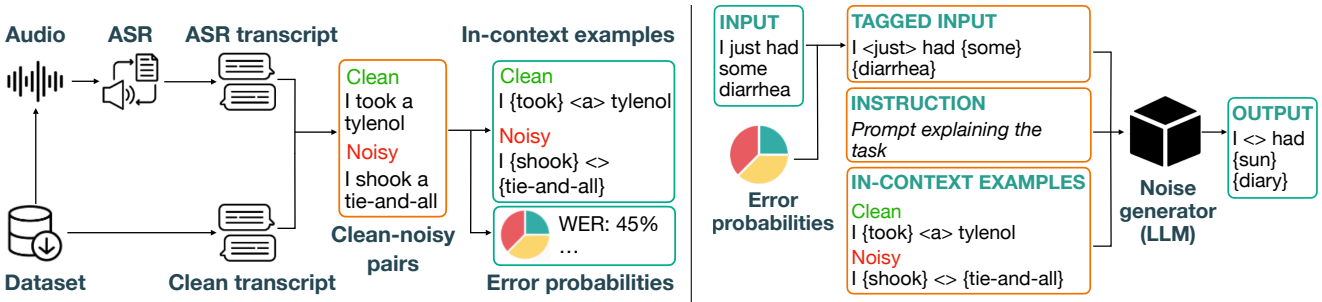
- We utilize the in-context learning ability of LLMs to create synthetic errors that closely resemble those present in ASR transcriptions. This method is used as a data augmentation strategy to improve the strength of dialogue summarization models in situations where audio recordings are limited.
- We introduce an error tagging syntax designed to accurately match the error patterns of real ASR transcriptions in the generated synthetic dialogues. This syntax provides detailed control over the type and amount of errors added, ensuring that the synthetic data closely aligns with real-world ASR errors.

2 Related Work

Dialogue Summarization Summarizing human conversations or speech has been a longstanding challenge that has received significant attention over the years (Hori and Furui 2003; Liu and Hakkani-Tür 2011). An important subset of these endeavors involves ASR-based summarization, where automatic speech recognition systems are used to transcribe spoken dialogues before summarization (Zhong et al. 2022). Recently, this challenge has extended to clinical applications, including the summarization of patient-doctor conversations (Krishna et al. 2021), utilizing few-shot (Le-Duc et al. 2024) or fine-tuned (Schlegel et al. 2023) LLMs as tools for text comprehension.

ASR Error Correction ASR models often produce erroneous dialogue transcripts, especially in noisy environments, degrading downstream tasks like summarization. Recent advancements in LLMs have shown promise in addressing these ASR errors. For instance, Yang et al. (2023) investigated prompting techniques in correcting ASR errors. Radhakrishnan et al. (2023) introduced Whispering LLaMA, a cross-modal generative error correction framework that fuses acoustic information with linguistic representations. Bai et al. (2024) proposed a seed-based method that enhances ASR outputs by integrating linguistic context during correction, effectively reducing error propagation in downstream tasks. Hu et al. (2024) introduced a method that leverages the inherent noise in audio signals to generate robust language embeddings. To provide a standard benchmark for evaluating ASR error correction methods, Chen et al. (2024) introduced the Hyporadise dataset. These techniques are unsuitable for the medical domain as fine-tuning denoising models is often impractical due to the lack of public medical dialogue audio recordings (Nanayakkara et al. 2022). Additionally, prompting without fine-tuning is only effective for large models (Yang et al. 2023), which may be impractical, due to privacy concerns when using public APIs or resource constraints when hosting locally.

Data Augmentation Data augmentation (DA) is crucial for enhancing model performance by diversifying training examples. Traditional DA methods in NLP, such as paraphrasing (Sharma et al. 2023; Li et al. 2023), back-translation (Sugiyama and Yoshinaga 2019), and noise injection (Wang et al. 2018), have been effective in some scenarios, but may not fully capture real-world variations. Recent LLM advancements have revolutionized DA by using models like GPT-3 and GPT-4 to generate high-quality synthetic data. For instance, Chintagunta et al. (2021) use in-context learning with LLMs to create synthetic medical dialogue



(a) Pre-processing stage. (i) Human-annotated clean dialogue transcripts are paired with ASR counterparts to compose in-context examples. (ii) Distribution of word-errors is inferred.

(b) Noisy data generation. Inferred word-error distribution guides error tagging in the input, which, along with instruction prompt and examples, is passed to the noise generator.

Figure 1: Overview of our MEDSAGE pipeline. In-context examples are created, and the ASR model's error profile is inferred. The error profile, in-context examples, and input dialogues are then processed by the LLM to generate noisy dialogues.

summaries, while Dialogic (Li et al. 2022) employs LLMs to generate annotated dialogues with automatic verification. Similarly, Liu et al. (2024) leverage topic-focused summarization and domain adaptation with LLMs to generate personalized medical dialogues. DA is also used to improve robustness against adversarial examples, as used by Liu et al. (2020), who apply adversarial training to make sentiment analysis models more resilient to noise and variations. Our method differs by addressing the challenges of augmenting medical dialogue data where ASR transcripts are scarce. Instead of heuristic corruptions that fail to mimic ASR errors accurately, we use LLMs to generate synthetic dialogues that better replicate ASR errors, ensuring augmented training data aligns with real-world test conditions.

3 MEDSAGE

We use LLMs to generate realistic ASR noise. On a high level, we pair human-transcribed sentences with their ASR counterparts as in-context examples to an LLM, which we instruct to generate noisy sentences (Section 3.1). These examples convey the qualitative aspects of the ASR errors such as phonetic confusions. To also ensure that the errors in the generated synthetic dialogues quantitatively match that of the ASR transcriptions, we propose a controlled generation strategy (Section 3.2). Our strategy involves introducing a tagging syntax that instructs the LLM on the specific word locations to corrupt with specific types of errors. This enables the generation of synthetic dialogues with any arbitrary error distribution that mirrors the characteristics observed in real ASR transcriptions. An overview of our MEDSAGE pipeline is given in Figure 1.

3.1 Error Generation using In-context Learning

To generate synthetic noisy dialogues, we first use a specialised system prompt to instruct the LLMs about the error generation task. Later we form the in-context examples by pairing sentences from human-transcribed clean dialogues with their ASR-generated counterparts. These examples are subsequently provided to the LLMs to convey the characteristics of ASR errors that we are trying to mimic. An example query that illustrates our structure when prompting the noise-generating LLMs is displayed in the following box.

In-Context Learning for ASR Noise Generation

System Prompt:
You are an AI assistant tasked with simulating errors similar to those made by Automated Speech Recognition (ASR) systems. You will be given sentences with the type of errors to be made indicated by tags. Corrupt the sentences based on `[explanation of the tagging system]`.

In-Context Examples

Input: I {took a Tylenol} `(human-transcribed)`
Response: I {shook tie-and-all} `(ASR)`

Input: I just had some {diarrhea} for the last three days `(human-transcribed)`
Response: I just had some {diary} for the last three days `(ASR)`

Input: yeah now i mean have you any have you noticed any kind of {white spots} on the back of your back of your {throat} or redness
Response:

As seen from the example, the clean sentences are given as inputs and the responses are expected to contain erroneous versions as if they were obtained through ASR models. The curly braces surrounding certain substrings are a part of our error tagging system, which is detailed later in Section 3.2, and they denote which words or substrings are transcribed wrongly by the ASR models and how. For instance, in the example provided, the word "Tylenol" is wrongly transcribed as "tie-and-all", highlighting the common issue of ASR systems confusing tokens with similar

pronunciation. Lastly, the input at the end contains the actual clean sentence that we aim to corrupt. The words to be corrupted are also indicated to the LLM through the use of curly braces.

3.2 Controlled Generation

For data augmentation to be effective in improving performance on noisy test data, the generated synthetic data must closely follow the error patterns found in real-world ASR outputs. However, solely relying on in-context learning with a few examples often fails to capture the nuanced distribution of errors produced by ASR models. To overcome this limitation, we propose a controlled noise injection mechanism. This is to, ensuring the synthetic data aligns with the error profile of the target ASR model. This approach relies on a detailed analysis of the types and frequencies of errors in ASR transcripts. Using these insights, we then guide the noise generation process by conditioning the LLM with an error tagging syntax.

ASR Error Profiling: To quantitatively represent the error profile of ASR models, we focus on three primary error types: insertion, deletion, and substitution, which collectively contribute to the calculation of Word Error Rate (WER). To estimate the distribution of these errors, we first transcribe a set of medical conversation dialogues using the ASR models. The transcriptions are then aligned with human-annotated ground truth using the Wagner-Fisher algorithm (Wagner and Fischer 1974). This alignment allows us to and quantify the occurrences of each specific error type produced by the ASR model.

We estimate the probability distribution of errors as follows. Let c_i denote the event that the word at index i is corrupted. The probability $p(c_i)$ of a word being corrupted is defined as the WER of the ASR model i.e., $p(c_i)$ = WER. Given that a word is corrupted, the probability of a specific error type e_t can be expressed as $p(e_t|c_i)$, where $e_t \in \{$*insertion, deletion, substitution*$\}$. These conditional probabilities represent the distribution of different error types observed in the ASR model's output (See Section 5.2). We use this error distribution to formulate our tagging system.

Tagging System We develop a tagging system to instruct the LLM on the specific word-level corruptions to perform to be able to have more control on the WER of generated noise and the distribution of error types. Specifically, we employ the following tags to indicate the error-type that the model should simulate at the word level.

- Words enclosed in curly brackets {} should be replaced with phonetically similar words. For instance, {`wheezy`} might be replaced with {`weesy`}, and {`Tylenol`} could be changed to {`tie-and-all`}.
- The tag `(INSERTION)` indicates where a new word should be inserted. These new words should be general in nature and should not introduce new domain-specific terminology such as drug names or symptoms.
- We do not specify any tags for deletion; instead, words that need to be deleted are simply removed from the text.

Algorithm 1: Generating Synthetic Noisy Dialogues Using LLMs

Input: Clean transcripts T_{cln}, in-context example pairs E_{in}, LLM
Output: Synthetic noisy dialogues T_{syn}

# Create in-context examples
for each pair $(t_{\text{cln}}, t_{\text{ASR}})$ in E_{in} **do**
    Compute word-locations and types of errors e_i, e_t
    $(t_{\text{cln}}, t_{\text{ASR}})' \leftarrow \text{insert\_tags}(t_{\text{cln}}, t_{\text{ASR}}, e_i, e_t)$
    Replace $(t_{\text{cln}}, t_{\text{ASR}})$ with $(t_{\text{cln}}, t_{\text{ASR}})'$ in E_{in}
end for

Initialize $T_{\text{syn}} \leftarrow \emptyset$

# Prompt LLM with in-context examples
for each input dialogue t_{cln} in T_{cln} **do**
    Sample error indexes $e_i \sim p(c_i)$
    Sample error types $e_t \sim p(e_t \mid c_i = e_i)$
    $t'_{\text{cln}} \leftarrow \text{insert\_tags}(t_{\text{cln}}, e_i, e_t)$
    $t_{\text{syn}} \leftarrow \text{LLM}(\text{prompt}, t'_{\text{cln}})$
    $T_{\text{syn}} \leftarrow T_{\text{syn}} \cup t_{\text{syn}}$
end for

This meaning of the tagging system is conveyed to the the model both through the system prompt and in-context examples. To decorate the in-context examples with error tags accordingly, we align the clean and noisy examples pairs using the Wagner-Fisher algorithm and determine the locations of word errors along with their types. During inference, these tags are randomly applied on the ground-truth transcripts based on the estimated error distribution of the target ASR model. The probability of tagging a word at index i with a specific error type is thus given by:

$$p(e_t|c_i) \cdot p(c_i) \tag{1}$$

which is the joint probability of a word being corrupted and the occurrence of a specific error type.

4 Experiment Settings

This section presents our experimental evaluation of the proposed method for generating synthetic noisy dialogue transcripts to improve the robustness of summarization models against ASR errors.

ASR models: We used the *Whisper tiny/ large*(Radford et al. 2023), and *Wav2vec2-base* (Baevski et al. 2020) ASR models to generate transcriptions of medical dialogues.

Large Language Models: We use *Llama-3-8B* (Dubey et al. 2024) and *Mistral-7B* (Jiang et al. 2023) both for generating synthetic dialogues and summarization, while *Gemma-7B* (Team et al. 2024) is only used for summarization.

Datasets: For testing our approach, we use the *Primock57* dataset, which comprises audio recordings of 57 enacted doctor-patient dialogues and their text transcripts written by human annotators. For experiments involving fine-tuning we

use the *NoteChat-1000* dataset (Wang et al. 2024) for training, which includes 1000 synthetic doctor-patient dialogue transcripts generated by multiple LLMs in a cooperative roleplay setting, conditioned on clinical notes.

Evaluation Metrics: We utilized three types of evaluation metrics to assess the performance of the summarization models. For lexical similarity, we used the ROUGE metrics (Lin 2004), specifically focusing on ROUGE-L, which measures the longest common subsequence overlap between generated and reference summaries. To capture the semantic similarity, we employed BERTScore (Zhang et al. 2020), which uses embeddings from pre-trained BERT models to compare the contextual meaning of the texts. Additionally, recognizing the importance of accurately identifying medical terminology in summaries, we assessed the overlap of domain-specific named entities (DSEs) using the F1 score, which combines both precision and recall, based on entities extracted through named entity recognition (NER).

5 Preliminary Experiments

In this section, we present preliminary experiments that lay the foundation for our work. First, we conduct a motivating study that verifies how ASR errors impair the performance of a downstream medical dialogue summarization task. Subsequently, we perform an analysis that suggests different ASR models exhibit distinct error profiles.

5.1 ASR noise harms medical report quality

Our findings, as displayed in Table 1, reveal that ASR noise can significantly degrade the quality of the generated summaries, especially when using small ASR models. Specifically, using Wav2vec2-base generated transcripts instead of human annotated ones lead to a noticeable reduction in the domain-specific entity F1 score of 23% (22.11 $\rightarrow$ 16.99). While using larger models like Whisper-large exhibits comparable downstream task performance as using ground truth dialogues, deployment area of such large-scale models is limited due to the computational resource requirements they demand. Such a difference in the DSE overlap measure indicates a loss in accurately capturing critical medical entities. The ROUGE-L scores also decline, reflecting reduced textual overlap and coherence between the generated and reference summaries. Moreover, the BERT Scores drop, suggesting a decrease in the semantic similarity to the reference summaries. These results underscore the challenges of downstream tasks, such as dialogue summarisation, face, arising from erroneous ASR transcripts. We also explored applying few-shot denoising on ASR transcribed dialogues, which is denoted as "+ Denoising". However, the results show that this method does not recover the summarization performance, highlighting the limitations of traditional post-processing based noise reduction techniques in this context.

	Llama-3-8B			Mistral-7B		
Transcription	**F1**	**RougeL**	**Bert**	**F1**	**RougeL**	**Bert**
Wav2vec2 (ASR)	16.99	11.52	52.46	17.06	11.09	50.40
+ Denoising	15.20	10.19	51.49	20.62	11.03	51.51
Whisper-T (ASR)	19.86	11.65	52.73	21.85	11.91	51.97
+ Denoising	12.73	9.95	51.94	19.81	11.42	51.74
Whisper-L (ASR)	21.42	13.67	**54.23**	24.06	**13.27**	**53.29**
+ Denoising	18.64	12.57	53.15	23.12	12.67	52.50
Ground Truth	**22.11**	**14.07**	53.80	**24.53**	13.25	52.85

Table 1: Effect of ASR errors on summarization quality. *"Ground Truth"* refers to using the ground-truth ASR transcripts for upper-bound performance. + *Denoising* indicates ASR transcripts cleaned by a denoising LLM before summarization.

5.2 Different ASR models exhibit different error profiles

We analyzed the errors made by the ASR models on Primock57 audio samples. *Whisper-large* transcription results in 25% WER. Both the *Whisper-tiny* and *Wav2vec2-base* models had similar WER scores, with the former achieving 44% and the latter 45%. Moreover, the breakdown of error types associated with each ASR model is displayed in Figure 2. The differing noise profiles suggest that to accurately mimic the properties of ASR transcriptions, the synthetic dialogue generation process must be controllable and adjustable with respect to the error profile of the target ASR model.

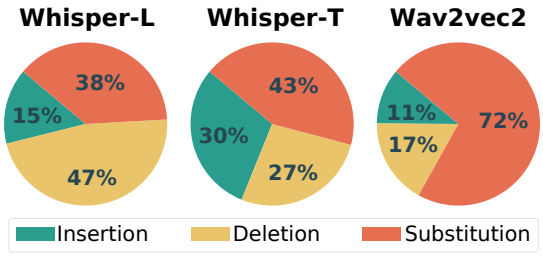


Figure 2: ASR errors of different models. The breakdown shows the different types of errors made by the model.

6 Main Experiments

Building on the insights from our preliminary experiments, we present our main experimental findings in this section.

6.1 Data augmentation using synthetic dialogues improves robustness against ASR errors

To assess the impact of MEDSAGE, we first investigate the effectiveness of LLM-generated synthetic dialogues in building robustness against ASR errors through data augmentation. In this experiment, we begin by augmenting the training set of the note chat-1000 dataset using synthetic dialogues. Later, we fine-tune summarization models through LoRa (Low-Rank Adaptation) adapters (Hu et al. 2021) on augmented datasets that include a mix of clean ground-truth dialogues and our synthetic noisy dialogues. Then we compare the summarization performance of these fine-tuned models against three baselines on the ASR-transcribed Primock57 audio recordings, which comprise our test set. The three baselines we use for evaluation are:

	Llama-3-8B			Mistral-7B			Gemma-7B		
	F1	**RougeL**	**Bert**	**F1**	**RougeL**	**Bert**	**F1**	**RougeL**	**Bert**
Zero-shot	19.9	11.8	**52.9**	20.8	11.8	52.4	21.3	12.7	51.0
+ Denoising	15.3 (-23.0%)	10.6 (-10.9%)	51.9 (-1.7%)	17.9 (-14.0%)	10.8 (-9.0%)	51.3 (-2.0%)	19.0 (-10.9%)	12.6 (-0.8%)	51.1 (+0.3%)
FT on clean	20.4 (+2.6%)	12.7 (+7.6%)	52.4 (-1.0%)	16.8 (-19.3%)	8.1 (-31.7%)	49.5 (-5.5%)	21.2 (-0.4%)	13.2 (+3.6%)	51.3 (+0.7%)
+ Denoising	19.3 (-2.8%)	11.5 (-3.0%)	51.4 (-2.8%)	14.9 (-28.3%)	7.5 (-36.6%)	48.7 (-6.9%)	19.1 (-10.4%)	12.5 (-1.9%)	50.6 (-0.7%)
MEDSAGE									
FT-aug (1x)	20.9 (+5.2%)	12.3 (+3.8%)	52.0 (-1.6%)	23.9 (+15.0%)	12.3 (+4.5%)	52.7 (+0.6%)	22.8 (+7.2%)	13.3 (+4.7%)	51.4 (+0.9%)
FT-aug (2x)	22.8 (+14.8%)	13.1 (+10.5%)	51.4 (-2.8%)	24.2 (+16.4%)	13.3 (+12.4%)	53.1 (+1.4%)	22.5 (+5.5%)	12.7 (0.0%)	51.5 (+1.0%)
FT-aug (3x)	22.5 (+13.1%)	12.8 (+7.9%)	51.9 (-1.8%)	22.1 (+6.5%)	12.9 (+9.5%)	52.7 (+0.6%)	21.0 (-1.6%)	13.3 (+4.1%)	51.3 (+0.7%)
FT-aug (Best)	**22.8 (+14.8%)**	**13.1 (+10.5%)**	52.0 (-1.6%)	**24.2 (+16.4%)**	**13.3 (+12.4%)**	**53.1 (+1.4%)**	**22.8 (+7.2%)**	**13.3 (+4.7%)**	**51.5 (+1.0%)**

Table 2: Summarization quality of Llama-3-8B, Mistral-7B, and Gemma-7B across training settings. *FT* and *aug* represent *fine-tuning* and *data augmentation*, respectively. In the *aug* setting, synthetic dialogues per ground truth are indicated by x multiples. Relative improvements (%) over Zero-shot are shown in parentheses.

- **Zero-shot**: Pre-trained LLMs are prompted to replicate medical notes written by doctors based on input dialogues.
- **FT on clean**: LLMs are fine-tuned only on clean transcripts.
- **Denoising**: The ASR transcripts are cleaned by LLama-3-8B model before summarization.

As shown in Table 2, the results indicate that the inclusion of synthetic noisy dialogues in the training set considerably improves the robustness of the summarization models, resulting in up to 16.4% improvement in F1. In contrast, *FT on clean* baseline only marginally improves or even degrades performance depending on the summarization model architecture. This suggests that while fine-tuning on clean dialogues alone can benefit summarization models by adapting them to the medical domain, it does not consistently enhance their ability to handle ASR noise. Moreover, denoising consistently harmed summarization performance across all experiments. This can be attributed to the two aforementioned key factors: (1) denoising models below a certain parameter size (such as LLama-8B) are ineffective at cleaning noise, and (2) using prompting techniques without fine-tuning introduces a risk of hallucinations, potentially injecting irrelevant medical entities into the dialogue.

6.2 LLM-produced synthetic errors are realistic

To verify that the improvements result from accurate ASR noise simulation, we must assess whether the synthetic dialogues from our MEDSAGE method exhibit error characteristics similar to actual ASR transcriptions. This involves two key comparisons.

Qualitative Comparison To illustrate that synthetic dialogues produced by our MEDSAGE method properly mimics the errors characteristics that real ASR transcriptions exhibit, we present randomly selected snippets from doctor-patient conversations in Figure 3. The commonalities between the nature of errors present in both utterances, such as phonetic confusions underscore the accuracy of the noise injections. For instance, the words "white spots" are confused with "whish spits". This qualitative evidence supports our quantitative findings, demonstrating that LLM-generated synthetic noise can replicate the nuanced flaws typically seen in ASR outputs.

Sentence
(Ground truth) Doctor: yeah now i mean have you any have you noticed any kind of white spots on the back of your back of your throat or redness **(Whisper ASR)** Doctor: yeah now i mean have you any if you know the new chrome white spots on the back ports youll back of your throws or readiness
(Ground truth) Doctor: yeah now i mean have you any have you noticed any kind of white spots on the back of your back of your throat or redness **MEDSAGE** Doctor: yeah now i mean have you any do ya notice any kind kinda whish spits on the back of your bak o yer throt or reddness

Figure 3: Comparison of an ASR transcription and its LLM-generated counterpart produced by MEDSAGE. The words highlighted in yellow are errors at common word indexes among both transcripts, while those that are highlighted in red and green indicate unique errors of ASR and MEDSAGE.

Quantitative Comparison To further substantiate the realism of LLM-generated synthetic noise, we conduct quantitative comparisons between the errors in synthetic dialogues and those in ASR-transcribed dialogues using the aforementioned metrics. First, we compute pairwise similarities, w.r.t. three evaluation metrics, among ASR transcripts and

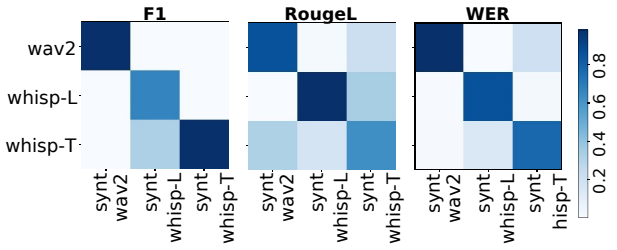


Figure 4: Similarities between ASR-generated transcriptions (rows) and LLM-generated synthetic transcriptions (columns) with respect to F1, Rouge-L, and WER metrics. The highest similarities are observed on the diagonals indicating overlap between corresponding ASR- and LLM-generated transcriptions.

Source Transcripts	Llama-3-8B Summarization			
	ASR	Random	LLM-generated (MEDSAGE)	
			Llama-3-8B	Mistral-7B
F1				
Wav2vec2-base	16.99	11.24	**17.35**	16.54
Whisper-tiny	19.86	12.83	**18.02**	17.91
Whisper-large	21.42	18.84	20.35	**22.02**
RougeL				
Wav2vec2-base	11.52	8.50	**11.47**	11.18
Whisper-tiny	11.65	9.88	**11.92**	10.91
Whisper-large	13.67	11.98	12.52	**12.70**
Bert				
Wav2vec2-base	52.46	49.59	52.87	**52.71**
Whisper-tiny	52.73	50.80	**53.48**	51.57
Whisper-large	54.23	52.07	**53.70**	53.59

Source Transcripts	Mistral-7B Summarization			
	ASR	Random	LLM-generated (MEDSAGE)	
			Llama-3-8B	Mistral-7B
F1				
Wav2vec2-base	17.06	11.73	**15.95**	15.05
Whisper-tiny	18.04	15.38	18.63	**18.10**
Whisper-large	24.06	22.28	**23.82**	21.40
RougeL				
Wav2vec2-base	11.09	8.32	**11.70**	10.34
Whisper-tiny	11.91	9.93	**12.60**	10.51
Whisper-large	13.27	11.82	**13.47**	11.39
Bert				
Wav2vec2-base	50.40	44.61	51.78	**49.56**
Whisper-tiny	51.97	48.17	**52.47**	48.67
Whisper-large	53.29	50.09	**53.64**	49.94

Table 3: Summarization quality for ASR and LLM-generated transcripts. *Random* denotes augmentation using random insertions, deletions, and substitutions. Blue ASR columns serve as reference, with bold values highlighting synthetic dialogues most similar to real ASR outputs.

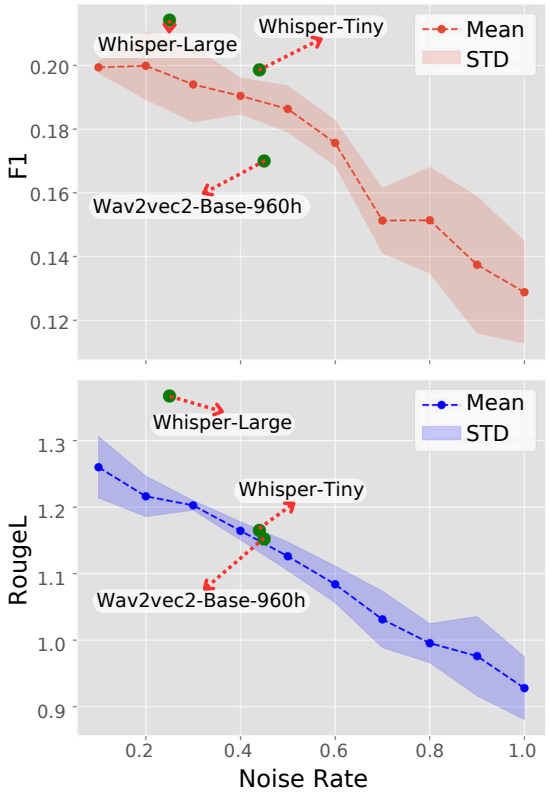


Figure 5: Change in the downstream summarization quality w.r.t. F1 and RougeL metrics as the rate of error tags used to generate synthetic dialogues increases. In-context examples were taken from Whisper-tiny transcribed dialogues. Green dots mark the scores of real ASR transcriptions.

synthetic dialogues. As shown in the similarity matrices in Figure 4, the diagonal entries contain the highest similarity scores, indicating that the greatest overlap is between ASR transcripts and the corresponding synthetic dialogues generated to mimic them. This pattern suggests that the LLM-generated synthetic dialogues are successful in accurately replicating the specific error characteristics of the ASR transcripts they were designed to imitate. Additionally, we provide an indirect comparison based on the summarization qualities each dialogue transcript yields. In this experiment, we also include dialogues corrupted with random deletions as well as insertions and substitutions from the NLTK words corpus (Bird, Klein, and Loper 2009) and name this baseline as *Random*. As shown in Table 3, the summarization qualities of LLM-generated synthetic dialogues closely align with those of the ASR transcripts, which are indicated in blue as reference values. This further supports the effectiveness of our method in producing realistic synthetic data.

6.3 Error tagging system enables controllable generation

Lastly, we validate that our method is capable of adjusting the injected noise rate through the use of corruption tags, as detailed in the methodology section. To this end, we insert corruption tags at various rates to simulate increasing amounts of noise. We then record the resulting transcription quality and summarization performance for each noise level. As depicted in Figure 5, we observe a clear trend: as the noise rate increases, the quality of the generated summaries decreases. This is evidenced by a gradual drop in the summarization quality quantified by the metric scores. This demonstrates the controllability of our noise generation method, which is essential for adapting the synthetic noise generation to errors made by various different ASR models.

7 Conclusion

This study highlights the critical role of Automated Speech Recognition (ASR) technology in transcribing medical dialogues and the consequent impact of ASR errors on downstream summarization tasks. Recognizing the challenges posed by limited availability of supervised data, we propose a novel data augmentation approach using large language models (LLMs) to generate synthetic noisy dialogues that mimic real-world ASR errors. Our findings demonstrate that LLMs can effectively model real-world ASR errors, and augmenting summarization models with these noisy transcripts leads to significant performance improvements.

Acknowledgements

Viktor is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Bai, Y.; Chen, J.; Chen, J.; Chen, W.; Chen, Z.; Ding, C.; Dong, L.; Dong, Q.; Du, Y.; Gao, K.; et al. 2024. Seed-ASR: Understanding Diverse Speech and Contexts with LLM-based Speech Recognition. *arXiv preprint arXiv:2407.04675*.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, C.; et al. 2024. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36.

Chintagunta, B.; Katariya, N.; Amatriain, X.; and Kannan, A. 2021. Medically Aware GPT-3 as a Data Generator for Medical Dialogue Summarization. In Shivade, C.; Gangadharaiah, R.; Gella, S.; Konam, S.; Yuan, S.; Zhang, Y.; Bhatia, P.; and Wallace, B., eds., *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, 66–76. Online: Association for Computational Linguistics.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Everson, K.; Gu, Y.; Yang, H.; Shivakumar, P. G.; Lin, G.-T.; Kolehmainen, J.; Bulyko, I.; Gandhe, A.; Ghosh, S.; Hamza, W.; et al. 2024. Towards ASR robust spoken language understanding through in-context learning with word confusion networks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12856–12860. IEEE.

Fabbri, A. R.; Han, S.; Li, H.; Li, H.; Ghazvininejad, M.; Joty, S.; Radev, D.; and Mehdad, Y. 2021. Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 704–717.

Guo, J.; Wang, M.; Qiao, X.; Wei, D.; Shang, H.; Li, Z.; Yu, Z.; Li, Y.; Su, C.; Zhang, M.; et al. 2024. Ucorrect: An unsupervised framework for automatic speech recognition error correction. *arXiv preprint arXiv:2401.05689*.

Hodgson, T.; and Coiera, E. 2016. Risks and benefits of speech recognition for clinical documentation: a systematic review. *Journal of the american medical informatics association*, 23(e1): e169–e179.

Hori, C.; and Furui, S. 2003. A new approach to automatic speech summarization. *IEEE Transactions on Multimedia*, 5(3): 368–378.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, Y.; et al. 2024. Large Language Models are Efficient Learners of Noise-Robust Speech Recognition. *arXiv preprint arXiv:2401.10446*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Krishna, K.; Khosla, S.; Bigham, J.; and Lipton, Z. C. 2021. Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4958–4972. Online: Association for Computational Linguistics.

Le-Duc, K.; Nguyen, K.-N.; Vo-Dang, L.; and Hy, T.-S. 2024. Real-time Speech Summarization for Medical Conversations. *arXiv preprint arXiv:2406.15888*.

Li, H.; Wu, Y.; Schlegel, V.; Batista-Navarro, R. T.; Nguyen, T.-T.; Kashyap, A. R.; Zeng, X.-J.; Beck, D.; Winkler, S.; and Nenadic, G. 2023. Team: PULSAR at ProbSum 2023: PULSAR: Pre-training with extracted healthcare terms for summarising patients' problems and data augmentation with black-box large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, 503–509.

Li, J.; Deng, L.; Gong, Y.; and Haeb-Umbach, R. 2014. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4): 745–777.

Li, Z.; Chen, W.; Li, S.; Wang, H.; Qian, J.; and Yan, X. 2022. Controllable Dialogue Simulation with In-context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4330–4347.

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

Liu, X.; Cheng, H.; He, P.; Chen, W.; Wang, Y.; Poon, H.; and Gao, J. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Liu, Y.; and Hakkani-Tür, D. 2011. Speech summarization. *Spoken language understanding: Systems for extracting semantic information from speech*, 357–396.

Liu, Z.; Ng, A.; Lee, S.; Aw, A. T.; and Chen, N. F. 2019. Topic-aware pointer-generator networks for summarizing spoken conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 814–821. IEEE.

Liu, Z.; Salleh, S.; Krishnaswamy, P.; and Chen, N. 2024. Context Aggregation with Topic-focused Summarization for Personalized Medical Dialogue Generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 310–321.

Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919.

Nagar, A.; Liu, Y.; Liu, A. T.; Schlegel, V.; Dwivedi, V. P.; Kaliya-Perumal, A.-K.; Kalanchiam, G. P.; Tang, Y.; and Tan, R. T. 2024. uMedSum: A Unified Framework for Advancing Medical Abstractive Summarization. *arXiv preprint arXiv:2408.12095*.

Nanayakkara, G.; Wiratunga, N.; Corsar, D.; Martin, K.; and Wijekoon, A. 2022. Clinical dialogue transcription error correction using Seq2Seq models. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, 41–57. Springer.

Papadopoulos Korfiatis, A.; Moramarco, F.; Sarac, R.; and Savkov, A. 2022. PriMock57: A Dataset Of Primary Care Mock Consultations. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 588–598. Dublin, Ireland: Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Radhakrishnan, S.; Yang, C.-H.; Khan, S.; Kumar, R.; Kiani, N.; Gomez-Cabrero, D.; and Tegnér, J. 2023. Whispering LLaMA: A Cross-Modal Generative Error Correction Framework for Speech Recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10007–10016.

Schlegel, V.; Li, H.; Wu, Y.; Subramanian, A.; Nguyen, T.-T.; Kashyap, A. R.; Beck, D.; Zeng, X.; Batista-Navarro, R. T.; Winkler, S.; and Nenadic, G. 2023. PULSAR at MEDIQA-Sum 2023: Large Language Models Augmented by Synthetic Dialogue Convert Patient Dialogues to Medical Records. In *Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023)*. Springer Lecture Notes in Computer Science LNCS.

Sharma, S.; Joshi, A.; Zhao, Y.; Mukhija, N.; Bhathena, H.; Singh, P.; and Santhanam, S. 2023. When and how to paraphrase for named entity recognition? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7052–7087.

Sugiyama, A.; and Yoshinaga, N. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the fourth workshop on discourse in machine translation (DiscoMT 2019)*, 35–44.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Wagner, R. A.; and Fischer, M. J. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1): 168–173.

Wang, J.; Yao, Z.; Yang, Z.; Zhou, H.; Li, R.; Wang, X.; Xu, Y.; and Yu, H. 2024. NoteChat: A Dataset of Synthetic Doctor-Patient Conversations Conditioned on Clinical Notes. arXiv:2310.15959.

Wang, L.; Fazel-Zarandi, M.; Tiwari, A.; Matsoukas, S.; and Polymenakos, L. 2020. Data augmentation for training dialog models robust to speech recognition errors. *arXiv preprint arXiv:2006.05635*.

Wang, X.; Pham, H.; Dai, Z.; and Neubig, G. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. *arXiv preprint arXiv:1808.07512*.

Yang, C.-H. H.; et al. 2023. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE.

Yu, D.; and Deng, L. 2016. *Automatic speech recognition*, volume 1. Springer.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhong, M.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11765–11773.