

Evaluating the Safety Tolerances of Speech Recognition AI for Look-Alike/Sound-Alike (LASA) Clinical Nomenclature under Acoustic Stress

Lokesh Chakrabarti

0040330@gsa1.org.uk

Grammar School

Short Report

Keywords:

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11471673/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

This study evaluates how open-source automatic speech recognition (ASR) models handle critical medical terminology when exposed to background noise, simulating the conditions of clinical transcription software used in hospitals. Software developers typically evaluate ASR models using an unweighted average Word Error Rate (WER). However, in high-acuity healthcare environments, aggregate metrics can hide dangerous, localized errors, such as the incorrect substitution of Look-Alike/Sound-Alike (LASA) medication names [1]. Using an automated Python testing pipeline, OpenAI's Whisper-Tiny model [2] was tested against clinical handover audio files overlaid with three tiers of Gaussian white noise (0%, 10%, and 30% noise amplitude configurations). The experimental results demonstrate a non-linear performance collapse under severe acoustic stress. Rather than showing a gradual decrease in accuracy, the model introduced critical phonetic mutations at low noise levels (e.g., "humalog" transcribed as "humilog") and entered infinite autoregressive repetition loops under high noise (30%), causing localized error rates to exceed 1200%. These findings indicate that standard commercial evaluation metrics fail to isolate safety risks in clinical deployment, highlighting the need for specialized validation frameworks.

Introduction

The deployment of ambient artificial intelligence scribes to automatically document patient consultations and clinical handovers is increasing within emergency medicine. While these tools significantly reduce the administrative burden on healthcare professionals, standard software evaluation methods present major vulnerabilities when applied to clinical environments. Developers routinely rely on global Word Error Rate (WER) to verify model accuracy. [3, 4].

While a 95% global accuracy rate is acceptable for general consumer applications, an unweighted average introduces significant clinical risk if the errors occur on critical pharmacological terms. Look-Alike/Sound-Alike (LASA) medications—compounds that sound nearly identical but serve completely different clinical functions—represent a serious blind spot for standard ASR deployment [1]. A single character error can corrupt an electronic medical record, converting a standard medication order into a severe patient safety hazard. This study utilizes an automated software simulation to stress-test an open-source language model under varying background noise levels to map exactly where the transcription pipeline fails.

Methodology

A linear testing script was set up in Python so the experiment could easily be run multiple times:

1. Four high-risk medical phrases were chosen based on real clinical safety guidelines [1], focusing on crucial electrolytes (hyperkalemia), chemotherapy (vinblastine), diabetes treatment (humalog), and blood pressure management (clonidine).

2. These text phrases were converted into clean UK-accent audio files (.wav) using the Google Text-to-Speech (gTTS v2.2.4) library [5].
3. The NumPy (v1.26.4) library [6] was used to add random white noise directly onto the audio data at three tiers: 0% (Quiet Control), 10% (Busy Ward), and 30% (Resus Chaos).
4. The noisy audio files were fed into the OpenAI whisper-tiny model (2022 release) running locally on a computer [2].
5. The jiwer (v4.0.0) library was used to calculate the WER [7], and a custom override was programmed so that if the AI completely missed the target drug, or swapped it with its alternative, the script instantly flagged it as a CRITICAL SAFETY FAILURE, regardless of its WER.

Results

The experiment showed that as background noise increased, the accuracy of the speech recognition model dropped sharply and unpredictably. Instead of a steady decline, the model experienced different types of complete translation failures at higher noise levels.

The complete breakdown of the Word Error Rates (WER) and the exact transcription errors can be seen in Table 1 below.

Table 1
Notes & Footnotes:

Case ID / Target Word	0% Noise (Quiet Control)	10% Noise (Simulated Busy Ward)	30% Noise (Resus Chaos)
1. Hyperkalemia (Electrolyte Management)	WER: 22.58% Output: "hyperkalemia" Status: Target Word Transcribed Perfectly	WER: 67.74% Output: "hypercellemia" Status: Target Word Critical Failure (Word Mutated)	WER: 961.29% Output: Repetition Loop ¹ Status: Total System Collapse
2. Vinblastine (Oncology / Chemo)	WER: 26.67% Output: "vinblastine" Status: Target Word Transcribed Perfectly	WER: 36.67% Output: "rimblasting" Status: Target Word Critical Failure (Wrong Word Swapped)	WER: 1266.67% Output: Repetition Loop ² Status: Total System Collapse
3. Humalog (Endocrinology / Diabetes)	WER: 33.33% Output: "humilog" Status: Target Word Critical Failure (Vowel Mutation)	WER: 56.67% Output: "cumolog" Status: Target Word Critical Failure (Wrong Word Swapped)	WER: 1100.00% Output: Repetition Loop ³ Status: Total System Collapse
4. Clonidine (Cardiovascular)	WER: 13.33% Output: "clonodin" Status: Target Word Critical Failure (Vowel Omission)	WER: 66.67% Output: "condolences" Status: Target Word Critical Failure (Everyday Word Swapped)	WER: 1103.33% Output: Russian Loop ⁴ Status: Total System Collapse

- **WER:** Word Error Rate (higher means more errors).
- **Critical Failure:** Triggered automatically by the testing script whenever the target medication name was missing, misspelled, or replaced.
- ¹ Model entered an endless loop repeating the phrase: *"the jesus, the jesus..."*
- ² Model entered an endless loop repeating the phrase: *"he is going to come out..."*
- ³ Model entered an endless loop repeating the phrase: *"he's been stuck in the house..."*

- ⁴ Model mistook the heavy white noise for foreign speech tokens, entering an endless loop in Russian: "что он был..."

Discussion & Analysis

The experimental data show that the speech recognition model experienced three distinct types of technical failure as acoustic noise increased:

- **Minor Character Mutations (0% Noise Scaling):** Even in the quiet control environment, the model struggled to accurately decode specific medical suffixes. The transcription of *clonidine* as "clonodin" and *humalog* as "humilog" represents trailing vowel omissions. In an active healthcare system, these minor typos break automated database queries, preventing digital prescribing systems from cross-checking the drug.
- **Out-of-Vocabulary Conversions (10% Noise Scaling):** When the background noise level increased, the model's internal decoding confidence dropped. Instead of outputting the correct clinical terms, the software substituted phonetically similar words from everyday language. The most severe safety failure occurred when *clonidine* (a high-potency blood pressure medication) mutated into the word "condolences". This error deletes the active drug name from the patient record, blinding automated hospital software to potential medication conflicts.
- **Autoregressive Hallucination Loops (30% Noise Scaling):** At the 30% noise threshold, the target audio was completely obscured by static. Because Whisper is an autoregressive model that relies on its own previous text outputs to predict the next word, a complete loss of acoustic input caused a feedback loop. The model stopped processing the timeline of the audio file and began repeating random phrases indefinitely. This technical vulnerability also explains the cross-lingual error in the *clonidine* run, where the model mistook heavy white noise for foreign speech tokens and entered a repetitive Russian text loop ("что он был...").

Conclusion

This simulation indicates that standard, unoptimized speech AI software can experience major structural failures when operated in high-noise environments. Conventional software benchmarks like global WER act as a lagging metric for safety, as high overall scores can easily mask a single, dangerous clinical word substitution. Future medical ASR development should move away from flat average metrics and integrate specialized semantic check layers that prioritize clinical terminology over conversational fluency.

Project Limitations & Future Directions

- **Acoustic Environment:** This script used continuous Gaussian white noise. Real emergency departments feature sudden, unpredictable sound impulses (such as medical alarms, shouting, and equipment noise). Future testing will replace uniform white noise with field recordings of active clinical environments.
- **Synthetic Speech Bias:** The use of mathematically consistent gTTS assets [5] excludes human speech variations like regional accents, stuttering, speed changes, and stress-induced vocal shifts.
- **Model Scale Limitations:** This study focused exclusively on a small, 39-million-parameter model (Whisper-Tiny). Larger model sizes or specialized clinical neural networks may offer greater resistance to hallucination loops, which requires comparative testing across different model parameters.
- **Dataset Size:** This pipeline functioned as a proof-of-concept validation framework limited to a four-term testing matrix. Expanding the testing dictionary to cover the complete ISMP high-alert index across a larger volume of trials is necessary to establish a statistically comprehensive safety profile.

Declarations

Ethics Approval:

- Not applicable. This study involves synthetic, computer-generated audio assets and code simulations. No human subjects, tissue samples, or patient clinical charts were utilized, exempting this work from institutional review board oversight.

Competing Interests:

- The author declares no competing financial, professional, or personal interests that could influence the execution or presentation of this study.

Funding:

The author received no external funding or financial backing for this research framework.

Data Availability Statement:

The computer simulation and experimental framework developed in this study are fully programmatic. The underlying datasets, audio simulations, and metrics are dynamically generated at runtime by the execution script.

Code Availability

Statement: The complete open-source Python source code, pipeline architecture, and dependency specifications are openly available in a public repository on GitHub at:

<https://github.com/LokeshChakrabarti/clinical-scribe-safety-audit>

References

1. Institute for Safe Medication Practices (ISMP) ISMP List of Confused Drug Names [Internet]. Horsham (PA): ISMP; 2023 [cited 2026 Oct 21]. Available from: https://www.ismp.org/system/files/resources/2023-10/ISMP_ConfusedDrugNames_2023.pdf
2. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356. 2022 [cited 2026 Oct 21]. Available from: <https://arxiv.org/abs/2212.04356>
3. Melnick ER, Natarajan S, Dinh VA et al (2024) Ambient artificial intelligence scribes in community emergency departments. Ann Emerg Med 84(4 Suppl 1):S173–S174. 10.1016/j.annemergmed.2024.07.411
4. Blease C, Kharko A, Annio G et al (2024) Generative AI and the future of clinical documentation: Opportunities and critical safety concerns. Lancet Digit Health 6(5):e290–e295. 10.1016/S2589-7500(24)00041-3
5. Durette PN pndurette/gTTS (Version 2.2.4) [Software]. GitHub. 2022 [cited 2026 Oct 21]. Available from: <https://github.com/pndurette/gTTS>
6. Harris CR, Millman KJ, van der Walt SJ et al (2020) Array Programming with NumPy. Nature 585(7825):357–362. 10.1038/s41586-020-2649-2
7. Vaessen N, JiWER Similarity measures for automatic speech recognition evaluation (Version 4.0.0) [Software]. GitHub. 2024 [cited 2026 Oct 21]. Available from: <https://github.com/jitsi/jiwer>