VOICE AND AUDIO

Automatic speech recognition in the exam room: word error rates, speaker diarization, and the errors that survive into the note

ABSTRACT

Ambient documentation products are judged almost exclusively by the fluency of the language model that drafts the final note. Yet every voice-first EMR sits downstream of an automatic speech recognition (ASR) and speaker-diarization pipeline that converts messy exam-room acoustics into text. Across a century of clinical speech-recognition evaluations, error rates vary by an order of magnitude between studio dictation and natural multi-speaker care: a 2018 audit of 217 dictated notes found 7.4 errors per 100 words before editing and 42.4% of physician-signed notes still containing errors (6.4% clinically significant); a 2025 systematic review of 29 AI transcription studies recorded word error rates (WER) from 8.7% in controlled dictation to over 50% in conversational care; and a 2024 home-healthcare evaluation measured median WERs of 39% to 73% across commercial and open-source engines, with consistent accuracy gaps for Black patients mirroring an earlier five-engine PNAS study (0.35 vs. 0.19 WER). Independent testing of medical-conversation diarization found word-level speaker-attribution errors up to 13.9%, while neural Whisper models fabricated entire unuttered phrases in roughly 1% of transcripts — disproportionately during pauses common in aphasia. Because standard WER weights a dropped negation identically to a dropped article, a voice-first record system must anchor drafted statements directly to source transcript utterances and preserve audio-to-note traceability.

Type Technical and evidence review **References** 25 **Reading time** 15 min **Last reviewed** November 2026

1 The silent upstream layer of every ambient scribe

Evaluations of ambient clinical documentation focus overwhelmingly on the large language model that turns a transcript into an HPI, physical exam and plan — a literature appraised in this library's reviews of ambient documentation and clinical language models. Far less clinical attention reaches the stage directly ahead of the language model: the automatic speech recognition (ASR) and speaker-diarization pipeline converting room acoustics into words and attributing those words to the clinician, the patient or an accompanying family member.[1,2]

Errors born in that first stage behave differently from ordinary LLM fabrications. A fluent language model rarely flags an implausible transcription; instead, it polishes a misheard drug name, a dropped "no" or a swapped speaker turn into grammatically seamless clinical prose that looks indistinguishable from an accurate record.[3,4]

7.4% → 0.3%	8.7% – >50%	0.35 vs. 0.19	1.8% – 13.9%
word error rate from raw dictation to signed note (42.4% of signed notes still had errors)	WER span across 29 clinical AI transcription studies from dictation to conversation	average WER across five major ASR engines for Black vs. White speakers	word-level speaker diarization error rate across four commercial ASR engines

This review synthesizes the peer-reviewed evidence on clinical ASR accuracy, speaker diarization, demographic and speech-impairment disparities, neural ASR hallucinations, and the record-architecture controls required when voice enters the legal health record.

2 Dictation versus exam-room conversation: why WER leaps

Clinical speech recognition did not begin with ambient AI; it matured over thirty years of single-speaker dictation, primarily in radiology and emergency medicine. In a systematic review of 122 studies from 1990 through 2018 published in *JAMIA*, Blackley and colleagues found that 39.3% of clinical speech-recognition evaluations were conducted exclusively in radiology, with reported word error rates spanning 7.4% to 38.7% and error-containing documents ranging from 4.8% to 71%.[5] Nine of the included workflow studies reported *longer* documentation times with speech recognition than with conventional transcription (from +13.4% to +200% in dictation plus correction time), and four found no significant change.[5,6]

To measure how many transcription mistakes survive human review, Zhou, Blackley, Kowalski and colleagues audited a stratified random sample of 217 dictated clinical notes (83 office notes, 75 discharge summaries, 59 operative notes) by 144 physicians at two health systems using commercial back-end speech recognition.[7] The raw speech-recognition error rate was 7.4 errors per 100 words (96.3% of raw notes contained errors), falling to 0.4% after professional medical transcriptionist editing (58.1% of notes) and 0.3% in the final physician-signed note — yet **42.4% of signed notes still contained errors**.[7] Deletions were the most common error class (34.7%), followed by insertions (27.0%). Crucially, 15.8% of raw errors, 26.9% of post-transcriptionist errors and 25.9% of errors surviving into signed notes involved clinical information, and 5.7%, 8.9% and 6.4% respectively were judged clinically significant.[7]

Moving from controlled single-speaker dictation to real-world multi-speaker clinical conversations degrades accuracy sharply. A July 2025 systematic review by Ng and colleagues covering 29 AI speech-recognition studies in *BMC Medical Informatics and Decision Making* documented word error rates from 0.087 (8.7%) in controlled dictation to more than 50% in conversational or multi-speaker clinical settings, with clinical-concept F1 scores spanning 0.416 to 0.856.[8] Table 1 contrasts controlled benchmarks against naturalistic clinical audio.

TABLE 1

Peer-reviewed benchmarks of clinical speech recognition and diarization across dictation, studio re-enactment and real patient encounters.

Study	Setting and audio corpus	Measured accuracy (WER / WDER)	Key operational takeaway
Zhou et al., 2018[7]	217 dictated notes, 144 physicians, 2 health systems	7.4% raw SR → 0.4% MT → 0.3% signed	42.4% of signed notes retained errors; 6.4% of surviving errors were clinically significant
Tran et al., 2023[1]	36 studio re-enacted primary-care visits (~134k words), 4 engines	WER 8.8%–10.5%; WDER 1.8%–13.9%	General engines beat medical-tuned engines on WDER; LOINC/SNOMED concept recall 0.49 on Amazon General
Zolnoori et al., 2024[9]	35 real home-healthcare nurse–patient encounters (mean age 67.8 y)	Mean WER 59% (AWS General), 62% (AWS Medical), 84% (Whisper)	Real-room acoustics and older/Black patient speech produced far higher error rates than studio tests
Ng et al., 2025[8]	Systematic review of 29 clinical AI ASR studies	WER 8.7% (controlled) to >50% (conversational)	Wide accuracy gap between single-speaker dictation and multi-speaker clinical dialogue

3 Speaker diarization: getting the words right for the wrong person

In an exam room, *who* said a sentence can invert its clinical meaning. If a patient says “my sister had a clot after surgery” and the clinician asks “any chest pain or shortness of breath today?”, misattributing the speaker turns can turn family history into personal history or turn a clinician’s screening question into a patient-reported symptom.[1,10]

Speaker diarization — partitioning an audio stream by speaker identity or clinical role — is measured alongside ASR via Word Diarization Error Rate (WDER), the fraction of words assigned to the wrong speaker.[10] In an independent evaluation published in the AMIA Annual Symposium Proceedings, Tran, Mangu, Tai-Seale, Lafata and Zheng tested four commercial cloud ASR systems (Amazon General, Amazon Medical Conversation, Google General and Google Medical Conversation) across 36 re-enacted primary-care conversations totaling roughly 134,000 words recorded in a quiet studio within three feet of a desktop condenser microphone.[1]

Even under those ideal studio conditions — two native English speakers of different sexes, lossless audio, no hallway noise, no overlapping family members — WER ranged from 8.8% (Google General) to 10.5% (Amazon Medical Conversation), and WDER spanned 1.8% (Amazon General) to 13.9% (Google Medical Conversation).[1] Contrary to marketing intuition, both *general-purpose* models outperformed their medical-conversation counterparts on diarization (Amazon General 1.8% vs. Amazon Medical 5.1%; Google General 6.3% vs. Google Medical 13.9%), and automated extraction of SNOMED CT and LOINC clinical concepts from Amazon General transcripts achieved 0.96 precision but only 0.49 recall.[1] Real clinic rooms — where a spouse answers on a patient’s behalf, a medical assistant steps in, or speakers talk over each other during a physical exam — present substantially harder diarization geometry than two studio actors reading turns.[11,12]

4 Demographic disparities, accents and speech impairments

Speech recognition accuracy is not uniform across patient populations. In a landmark 2020 study in the *Proceedings of the National Academy of Sciences*, Koenecke and colleagues evaluated five commercial ASR engines (Amazon, Apple, Google, IBM and Microsoft) on 19.8 hours of structured interviews across five US cities (42 White speakers and 73 Black speakers, matched by age and gender).[13] Every single engine exhibited substantial racial disparities, with an average WER of 0.35 for Black speakers compared with 0.19 for White speakers. Because the gap persisted identically on the subset of identical phrases spoken by Black and White speakers, the authors traced the disparity directly to acoustic-model underrepresentation of African American Vernacular English rather than vocabulary differences.[13]

Four years later, Zolnoori and colleagues tested whether those gaps persisted in actual clinical encounters.[9] Analyzing 860 utterances from 35 older adult home-healthcare patients (16 Black, 19 White; mean age 67.8 years) speaking with visiting nurses in New York City, they found dramatically higher error rates across all four tested engines: AWS General averaged 59% WER (median 39%), AWS Medical 62% (median 47%), Whisper 84% (median 73%) and Wav2Vec 99% (median 91%).[9] Every engine transcribed Black patients less accurately than White patients, with disparities pronounced in emotion, social and narrative language categories — the exact segments where clinicians gauge pain interference, mood and treatment adherence.[9]

Because HHS's Section 1557 nondiscrimination rule reaches covered practices using patient-care decision support tools — examined in this library's paper on algorithmic bias and Section 1557 — systematic transcription failure for non-standard dialects, older patients, regional accents or dysarthria is both a clinical safety issue and a health-equity liability.[14,15]

5 Neural ASR hallucinations and the limits of Word Error Rate

Older acoustic-plus-n-gram speech recognizers made phonetic mistakes: mishearing "metformin" as "met for men." Modern sequence-to-sequence transformer ASR models introduce a qualitatively different failure mode: *inventing fluent text during silence*.[16]

In "Careless Whisper," published at ACM FAccT 2024, Koenecke, Choi, Mei, Schellmann and Sloane audited OpenAI's Whisper on speech datasets including speakers with aphasia.[16] Roughly 1% of transcriptions contained entire hallucinated phrases or sentences absent from the underlying audio, and 38% of those hallucinations carried explicit harms — including fabricated references to violence, false associations or invented authority.[16] Logistic regression showed that hallucinations occurred disproportionately during longer non-vocal intervals (pauses and silences), meaning patients with aphasia, dyspnea, severe pain or hesitant speech face systematically higher fabrication risk during silent stretches of an exam.[16]

Why WER hides clinical danger

Standard Word Error Rate treats every substitution, deletion and insertion equally: `WER = (S + D + I) / N`. On a 1,000-word visit transcript, dropping “um,” “well,” and “actually” three times creates a 0.3% WER with zero clinical consequence. Dropping the single word “no” in “no new leg weakness,” substituting “fentanyl” for “fexofenadine,” or shifting “50 micrograms” to “50 milligrams” produces a 0.1% WER while threatening patient safety.[3,4,17]

Clinicians are aware of the risk even when they like the speed: in a survey of 1,731 speech-recognition users across two US health systems, Goss and colleagues reported that 78.8% were satisfied and 77.2% agreed speech recognition improved efficiency, yet 19.6% estimated that half or more of the errors they encountered were clinically significant.[18,19] Recent frameworks evaluating ambient voice technologies therefore argue against relying on global WER alone, recommending entity-level accuracy for medications, dosages, laterality, negation polarity and allergy status, plus downstream error-propagation audits through the LLM summarizer.[2,4,20]

LIMITATIONS OF CURRENT ASR-TO-LLM PROPAGATION STUDIES

Most published clinical ASR studies evaluated either standalone transcription engines (AWS, Google, Whisper, Dragon) or end-to-end commercial ambient scribe products without exposing the intermediate transcript.[1,8,9] Public, peer-reviewed measurements of how often a current-generation language model self-corrects an ASR substitution versus silently cementing it into the final clinical note remain sparse as of late 2026.[2,20]

6 Architectural requirements for a voice-first EMR

If upstream speech recognition will inevitably mishear words, swap speakers occasionally and misbehave during long pauses, a voice-first EMR cannot treat the transcript as a throwaway buffer. Five concrete engineering properties separate a verifiable voice workflow from an opaque scribe:[2,4,21]

1. **Sentence-level transcript anchoring.** Every AI-drafted sentence in the encounter note should link back to the exact timestamped transcript turn(s) that produced it, allowing the signing clinician to verify a surprising dosage, negation or attribution in one click without re-reading the entire transcript.[2,22]
2. **Voice-activity detection and silence gating before ASR.** Suppressing silence segments and non-speech room noise before decoding sharply reduces the long non-vocal windows that trigger transformer ASR hallucinations, while preserving natural pauses for patients with aphasia or chronic pain.[16]

3. **Structured reconciliation against chart state.** Never let raw spoken drug names or dosages flow unchecked into orders or medication lists: cross-check transcribed entities against RxNorm, the active medication list, allergies and prior MME/day baselines, surfacing discrepancies explicitly for clinician review.[3,23]
4. **Visible speaker-role boundaries with low-friction reassignment.** Expose clinician, patient and caregiver speaker labels cleanly in the live transcript and make role correction straightforward when a third party speaks for the patient.[1,12]
5. **Durable draft provenance and subgroup monitoring.** Record transcript capture, AI note drafting and clinician sign-off as separate, tamper-evident events (see write-path architecture and hash-chain records), and monitor clinician edit rates across acoustic environments and patient demographics so silent accuracy regressions surface early (see monitoring after go-live).[9,24,25]

The empirical lesson across thirty years of clinical speech recognition is consistent: voice capture saves effort only when proofreading and verification are engineered into the chart itself, rather than left to an exhausted clinician scanning prose at the end of clinic.

REFERENCES

Entries 1, 5 to 10, 13, 16, 18, 19 and 22 to 25 are peer-reviewed journal articles, systematic reviews and conference papers in JAMA Network Open, JAMIA, JAMIA Open, PNAS, BMC Medical Informatics and Decision Making, AMIA, Interspeech and ACM FAccT. Entries 2, 3, 4, 11, 12, 14, 15, 17, 20 and 21 include recent 2024–2026 evaluations, evaluation frameworks and preprints on ASR-LLM integration and clinical safety, noted as such where cited.

1. Tran BD, Mangu R, Tai-Seale M, Lafata JE, Zheng K. *Automatic Speech Recognition Performance for Digital Scribes: A Performance Comparison Between General-Purpose and Specialized Models Tuned for Patient-Clinician Conversations*. AMIA Annual Symposium Proceedings. 2023;2022:1072–1080; PMCID: PMC10148344. pmc.ncbi.nlm.nih.gov BENCHMARK
2. Automatic Speech Recognition in Healthcare in the Post-LLM Era. *PubMed Central Review*. 2026; PMCID: PMC13205700. Examines cascading error mitigation, clinical safety validation and subgroup fairness in ASR-LLM pipelines. pmc.ncbi.nlm.nih.gov REVIEW
3. Revisable Trailing Context in LLM-Grounded Clinical Speech Recognition. *OpenReview*. Examines how misheard medication names and garbled dosages propagate from ASR into generated notes and order entry. openreview.net PREPRINT
4. An Evaluation Framework for Ambient Digital Scribing Tools in Clinical Practice. *PubMed Central*. June 2025; PMCID: PMC12166074. pmc.ncbi.nlm.nih.gov FRAMEWORK

5. Blackley SV, Huynh J, Wang L, Korach Z, Zhou L. *Speech Recognition for Clinical Documentation From 1990 to 2018: A Systematic Review*. Journal of the American Medical Informatics Association. 2019;26(4):324–338. doi:10.1093/jamia/ocy179 SYSTEMATIC REVIEW

6. Hodgson T, Magrabi F, Coiera E. *Efficiency and Safety of Speech Recognition for Documentation in the Electronic Health Record*. Journal of the American Medical Informatics Association. 2017;24(6):1127–1133. doi:10.1093/jamia/ocx073 EXPERIMENTAL

7. Zhou L, Blackley SV, Kowalski L, Doan R, Acker WW, Landman AB, et al. *Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists*. JAMA Network Open. 2018;1(3):e180530. doi:10.1001/jamanetworkopen.2018.0530 CROSS-SECTIONAL

8. Ng JJW, Wang E, Zhou X, Zhou KX, Goh CXL, Sim GZN, et al. *Evaluating the Performance of Artificial Intelligence-Based Speech Recognition for Clinical Documentation: A Systematic Review*. BMC Medical Informatics and Decision Making. 2025;25(1):236. doi:10.1186/s12911-025-03061-0 SYSTEMATIC REVIEW

9. Zolnoori M, Vergez S, Xu Z, Esmaeili E, Zolnour A, Briggs KA, et al. *Decoding Disparities: Evaluating Automatic Speech Recognition System Performance in Transcribing Black and White Patient Verbal Communication With Nurses in Home Healthcare*. JAMIA Open. 2024;7(4):ooae130. doi:10.1093/jamiaopen/ooae130 OBSERVATIONAL

10. El Shafey L, Soltau H, Shafran I. *Joint Speech Recognition and Speaker Diarization via Sequence Transduction*. Proceedings of Interspeech 2019:396–400. Defines Word Diarization Error Rate (WDER). doi:10.21437/Interspeech.2019-1943 CONFERENCE PAPER

11. Mao HH, Li S, McAuley J, Cottrell GW. *Speech Recognition and Multi-Speaker Diarization of Long Conversations*. Proceedings of Interspeech 2020. isca-archive.org CONFERENCE PAPER

12. LLM-Based Speaker Diarization Correction: A Generalizable Approach. *arXiv preprint arXiv:2406.04927*. Revised March 2025. arxiv.org/abs/2406.04927 PREPRINT

13. Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, et al. *Racial Disparities in Automated Speech Recognition*. Proceedings of the National Academy of Sciences. 2020;117(14):7684–7689. doi:10.1073/pnas.1915768117 BENCHMARK

14. Accents Still Confuse AI: Systematic Errors in Speech Transcription and LLM-Based Remedies. September 2025. Examines accent-linked errors on medication names, dosages and patient instructions. researchgate.net PREPRINT

15. ASR Under the Stethoscope: Evaluating Biases in Clinical Speech Recognition Across Indian Languages. December 2025. Evaluates speaker-role, gender and code-mixed transcription disparities. researchgate.net PREPRINT

16. Koenecke A, Choi ASG, Mei KX, Schellmann H, Sloane M. *Careless Whisper: Speech-to-Text Hallucination Harms*. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24):1672–1681. doi:10.1145/3630106.3658996 CONFERENCE PAPER

17. Accuracy of Large Language Model Transcription of Simulated Clinical Encounters. *BMC Medical Informatics and Decision Making*. 2026; PMCID: PMC13088782. pmc.ncbi.nlm.nih.gov EXPERIMENTAL

18. Goss FR, Zhou L, Weiner SG. *Incidence of Speech Recognition Errors in the Emergency Department*. International Journal of Medical Informatics. 2016;93:70–73. doi:10.1016/j.ijmedinf.2016.05.005 CROSS-SECTIONAL

19. Goss FR, Blackley SV, Ortega CA, Kowalski LT, Landman AB, Lin CT, et al. *A Clinician Survey of Using Speech Recognition for Clinical Documentation in the Electronic Health Record*. International Journal of Medical Informatics. 2019;130:103938. doi:10.1016/j.ijmedinf.2019.07.017 SURVEY

20. Impact of Acoustic and Informational Noise on AI-Generated Clinical Summaries. October 2026. Evaluates downstream summary sensitivity to exam-room acoustic degradation. researchgate.net PREPRINT

21. Evaluating the Application of Artificial Intelligence and Ambient Listening in Clinical Documentation. *PubMed Central*. June 2025; PMCID: PMC12146405. pmc.ncbi.nlm.nih.gov REVIEW

22. Real-Time Multilingual Speech Recognition and Speaker Diarization System for Clinical Settings. *PubMed Central*. March 2024; PMCID: PMC11041969. pmc.ncbi.nlm.nih.gov TECHNICAL STUDY

23. Speech Recognition Technology and Documentation Efficiency. *JAMA Network Open*. 2025;8(3):e251526. doi:10.1001/jamanetworkopen.2025.1526 COHORT

24. Onitilo AA, Shour AR, Puthoff DS, Tanimu Y, Joseph A, Sheehan MT. *Evaluating the Adoption of Voice Recognition Technology for Real-Time Dictation in a Rural Healthcare System: A Retrospective Analysis of Dragon Medical One*. PLoS One. 2023;18(3):e0272545. doi:10.1371/journal.pone.0272545 COHORT

25. Understanding Clinician Edits to Ambient AI Draft Notes: A Feasibility Analysis Using Large Language Models. March 2026. researchgate.net PREPRINT

RELATED PAPERS

Ambient documentation and the clinician’s day

What randomized trials measured downstream once drafts reach the note.

A write-path architecture for an AI-native EMR

How drafts stay anchored to the recorded transcript until clinician sign-off.

Bias in clinical algorithms and Section 1557

How accuracy disparities across patient subgroups intersect with federal law.