

Medication Names Are a Safety-Critical Failure Mode in Clinical Speech Recognition: An Open Benchmark of On-Device and Cloud Systems

J. Grey Faulkenberry

grey.faulkenberry@emory.edu

Emory University School of Medicine

Research Article

Keywords: Speech recognition software, Medication errors, Documentation, Mobile applications, Benchmarking

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10804600/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Ambient AI scribes now entering clinical practice depend on speech recognition that transmits patient audio to cloud services, and no open benchmark exists for clinical conversation transcription. We built an open, reproducible benchmark and used it to characterize clinically consequential failure modes across deployment settings, and secondarily to determine whether on-device models reach cloud-level clinical term accuracy and whether multi-model fusion narrows any residual gap.

Methods

We evaluated 12 on-device ASR models and 5 cloud APIs on 400 simulated conversations from three public datasets (OSCE respiratory, $n = 272$; PriMock57 primary care, $n = 57$; Kazi et al. psychiatric, $n = 71$). WER was reported in standard (Whisper normalizer) and meaning-preserving regimes. Clinical term recall (CTR) used UMLS 2025AB spans via QuickUMLS with BCa bootstrap 95% CIs. ROVER and six advanced fusion algorithms were evaluated.

Results

Medication names were the highest-error clinical term category for most systems in both deployment modes; the highest-error drugs were misrecognized in 85% to 93% of on-device occurrences versus 3% to 30% in the cloud. A medical-dictation model cut wellbutrin errors to 5%, showing this deficit is addressable through domain adaptation. On-device clinical term recall was within 1–3 pp of the best cloud service per dataset; the standard-WER gap was 0.5–5.2 pp, narrowing to 0.7–3.3 pp under meaning-preserving normalization. ROVER gained ≤ 0.82 pp; no advanced algorithm exceeded naive voting by more than 0.4 pp.

Conclusions

Medication errors reflect look-alike/sound-alike confusability, and their persistence alongside near-parity recall shows that aggregate accuracy can mask clinically critical failure modes. Because this failure is independent of where transcription runs, any clinical ASR deployment requires explicit medication verification before note finalization. Within that constraint, on-device ASR is a clinically viable, lower-cost alternative that keeps patient audio local, and a single model suffices; fusion did not close the small residual gap.

Background

Physicians spend more than half their working hours on documentation, and administrative burden is a leading contributor to burnout.[1, 2] Ambient AI scribes, which passively transcribe encounters for note generation, have emerged as a promising intervention. Evaluations report reductions in documentation time and improved clinician satisfaction,[3, 4] though results are mixed: one large cohort study of a commercial ambient scribe found no significant time savings and worsened after-hours EHR use.[5] Regardless of downstream effectiveness, all such systems depend on the accuracy of automatic speech recognition (ASR), where reported word error rates range from below 10% for controlled dictation to over 50% for naturalistic clinical speech.[6]

Commercial cloud ASR services now offer medical-specific models, but these require transmitting patient audio to remote servers, and some ambient scribe services retain full audio recordings for weeks after the encounter.[7] Business Associate Agreements (BAAs) address HIPAA liability but do not eliminate the privacy risks of transmitting and storing protected health information with third parties, and recent analyses raise broader concerns about consent, transparency, and data governance.[7, 8] On-device ASR, which processes audio locally without network transmission or third-party storage, addresses these concerns. Open-source ASR models have improved rapidly, with several now suitable for mobile and edge deployment, yet no standardized benchmark compares them against commercial cloud APIs on clinical conversation data.

Existing evaluations of medical ASR have significant limitations. Industry benchmarks report WER on proprietary datasets that cannot be independently verified. Afonja et al. introduced a medical WER for accented clinical speech but relied on a commercial NER service to label medical entities, limiting reproducibility.[9] A recent single-center evaluation of large language model transcription of simulated encounters found that medical terminology errors accounted for 15% of word errors and 42% of semantically consequential errors, but it examined twelve encounters with three standardized patients, and its authors note that this limited clinical diversity restricts generalizability.[10] Ng et al.'s systematic review noted the absence of a standardized evaluation framework for clinical ASR,[6] and the socio-technical risks of clinical speech-to-text systems remain largely unaddressed.[7]

We present the first open, reproducible benchmark of ASR for medical conversations spanning both deployment modes, and use it to show that medication names are a safety-critical failure mode independent of where transcription runs: they are the highest-error clinical term category for most on-device models and remain the weakest cloud category on two of three datasets. On-device ASR otherwise reaches near-parity with commercial cloud services, so keeping patient audio on the device imposes no meaningful accuracy penalty. The benchmark comprises: (1) a standardized comparison of 12 on-device models and 5 cloud APIs across three public datasets totaling 400 clinical conversations; (2) a clinical term recall analysis using UMLS; (3) a per-term error analysis identifying medication names as the highest-error category for most models, with direct patient-safety implications; (4) an evaluation of ROVER fusion across pairwise and three-model combinations, showing a single model suffices; and (5) a cost analysis of cloud versus zero-marginal-cost on-device inference.

Methods

Datasets

We evaluated three public English-language clinical conversation datasets spanning different specialties and recording conditions.

The OSCE respiratory interview dataset[11] contains 272 simulated patient-physician interviews focused on respiratory cases, recorded in a controlled OSCE examination setting; audio and manually corrected transcripts are available under a CC0 1.0 public-domain dedication on figshare.

PriMock57[12] contains 57 mock primary care consultations recorded between medical professionals acting as doctor and patient. The dataset includes audio recordings, utterance-level transcripts, and consultation notes, released under a CC-BY 4.0 license.

The Kazi et al. psychiatric dataset[13] contains 71 enacted psychiatric consultations generated by pairs of students reading from clinical transcripts, released under a CC-BY 4.0 license. All three datasets consist of simulated or enacted encounters; no open-license dataset of real patient-clinician conversations currently exists (see Limitations).

On-Device Models

We evaluated 12 publicly available ASR models spanning four architectural families: encoder-decoder (Whisper distil-v3.5, Whisper turbo, Whisper base.en, Qwen3-ASR), transducer (parakeet-tdt-0.6b-v2, NeMo FastConformer, Nemotron, Zipformer), CTC (SenseVoice, SenseVoice without ITN, MedASR), and non-autoregressive (Paraformer-en). Model parameter counts ranged from 67M (Zipformer) to 809M (Whisper turbo). SenseVoice was evaluated with and without inverse text normalization (ITN), the only model in the pool to expose ITN as a runtime option; both settings are used in deployment. All on-device models were evaluated using their published int8-quantized builds, the form distributed for edge deployment; in two of these builds the upstream distribution leaves one component unquantized (the Zipformer decoder and the Qwen3-ASR convolutional frontend).

Eleven of the 12 on-device models are general-purpose; the twelfth, MedASR, is a medical-dictation model. All were run through the sherpa-onnx inference engine (v1.12.34), an open-source runtime that executes ONNX models on CPU without GPU acceleration for mobile and edge deployment. Benchmarks ran on a Lenovo ThinkPad P14s Gen 4 (AMD Ryzen 7 PRO 7840U, 64 GB RAM, Ubuntu 24.04 LTS); inference was faster than real-time for all models. The same 12 models were evaluated on all three datasets for consistent comparison.

Cloud APIs

Five commercial cloud ASR services were evaluated: Azure Speech (Batch Transcription v3.2), Google Cloud Speech (v1, medical_conversation model), Deepgram Nova-2 Medical, AssemblyAI (Universal-3 Pro, invoked with the system prompt “This is a doctor-patient clinical conversation.”), and AWS Transcribe Medical (Specialty=PRIMARYCARE, Type=CONVERSATION). Three of the five services (Google, Deepgram, AWS) used medical-specific models, Azure and AssemblyAI general-purpose models. Azure, Deepgram, AssemblyAI, and AWS Transcribe Medical were run on all 400 files; Google Cloud Speech on 396: four of the six longest psychiatric recordings had not completed within our 10-minute client-side wait and were not retried. Because each service exposes different clinical-configuration controls (Google and Deepgram offer medical model selection, AWS specialty and type parameters, AssemblyAI a free-text system prompt, and Azure none), these settings are not strictly equivalent across providers; the AssemblyAI clinical system prompt in particular has no exact analog in the other services and may modestly advantage or disadvantage it.

Text Normalization and WER Computation

Word error rate was computed under two normalization regimes. **Standard WER** applied the Whisper English text normalizer (whisper-normalizer v1.0), which performs case folding, punctuation removal, number-to-word conversion, and contraction expansion; this is the de facto pipeline used by the HuggingFace Open ASR Leaderboard[14] and MLPerf Inference ASR benchmarks, reported here for direct comparability. **Meaning-preserving WER** additionally applies a normalization layer (described below) that collapses orthographic variants and conversational backchannels; both regimes are reported for every model and dataset. Because we evaluated an exhaustive set of models and files rather than sampling from a population, we report descriptive performance metrics without hypothesis tests. To quantify file-level variability in clinical term recall we computed 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals over files (10,000 resamples). Resampling used common random numbers within each dataset (a fixed per-dataset seed), so the intervals are reproducible and any future model can be added without recomputing the others, while model-versus-model comparisons draw matched resamples. These intervals are reported alongside each recall estimate.

The OSCE dataset reference transcripts were distributed with apostrophes removed from contractions (e.g., “IM” for “I’m”), which interacted with the normalizer’s contraction expansion to inflate WER for all systems on this dataset. We corrected this by injecting apostrophes into the reference using a 44-token contraction dictionary, of which 39 were unambiguous (e.g., “dont” → “don’t”) and 5 were validated against the hypothesis texts as predominantly contraction forms (its, ill, id, lets, wed). Two ambiguous tokens (well, were) were left uncorrected because the bare-word interpretation dominated in the hypotheses. All reported OSCE WER values reflect this corrected reference; the PriMock57 and psychiatric references were unaffected.

Clinical Text Normalization

The Whisper normalizer is domain-agnostic: it does not collapse orthographic variants common in clinical text that do not change meaning, nor remove conversational backchannels that some services transcribe verbatim and others suppress. We added a meaning-preserving layer on the Whisper baseline, with rules in eight categories. (1) **British/American spelling**: a 148-entry dictionary (e.g., diarrhoea/diarrhea, oedema/edema, colour/color) plus suffix rules (-our/-or, -ise/-ize, -re/-er, -ae/-e, -oe/-e). (2) **Hyphenation**: hyphens replaced with spaces so that “x-ray”, “xray”, and “x ray” map identically. (3) **Open versus closed compounds**: a 42-entry dictionary (e.g., health care/healthcare, post operative/postoperative). (4) **Possessive eponyms**: common medical eponyms (Crohn, Alzheimer, Parkinson, and others) canonicalized so that “Crohn’s”, “Crohns”, and “Crohn disease” map identically; the possessive ’s is stripped before the Whisper normalizer expands it into “is”. (5) **Spaced acronyms**: 54 medical acronyms (CT, MRI, ECG, COPD, BID) joined when transcribed as spaced single letters (“c t scan” → “ct scan”). (6) **Honorifics**: abbreviated titles (Dr., Mr., Mrs., Ms.) expanded to spelled-out forms. (7) **Dosing units**: spelled-out units (milligrams, micrograms, milliliters) collapsed to abbreviations (mg, mcg, ml). (8) **Conversational backchannels**: a closed 21-token list (e.g., um, uh, er, ah, hmm, mm, mhm, uhuh, huh, hum, yeah, yep, yup, okay, ok, oh) stripped symmetrically from reference and hypothesis so that systems transcribing them verbatim are not penalized against systems that suppress them. We deliberately retained ambiguous tokens that can carry clinical meaning (e.g., “right” can refer to a side, “well” to health status, “no” and “yes” to clinical decisions). The full rule set (~ 530 lines, with self-tests) is in scripts/extended_normalizer.py; it is applied symmetrically, is idempotent, and runs over stored hypotheses without re-running ASR inference.

Recognizer Output Voting Error Reduction (ROVER)[15] combines hypotheses by aligning them at the word level and selecting the most frequent word at each position. Models with solo WER exceeding 30% on two or more datasets were excluded because weak models degraded fused output, yielding a pool of 9 competitive on-device models (the fusion runs predate the int8 consolidation and used the full-precision NeMo FastConformer build, which appears in no best-performing combination). We evaluated all 36 pairwise combinations with equal weighting on PriMock57 and the psychiatric dataset, and all 84 three-model combinations on PriMock57 using progressive pairwise fusion (equal weighting at each step). Exhaustive triple search on the psychiatric and OSCE datasets was infeasible in our pure-Python implementation ($O(n^2)$ per-file alignment); for OSCE we report only the targeted fusion of the two best solo on-device models (parakeet-tdt-0.6b-v2 + sensevoice). Complete pair and triple results are in Supplementary Tables S2 and S3. To test whether sophisticated algorithms exceed naive equal-weight voting, we evaluated six fusion variants on the pairing of the two best solo models (parakeet-tdt-0.6b-v2 + whisper-distil-v3.5) on PriMock57, the dataset with the largest pairwise fusion gain: confusion network combination with epsilon tuning;[16] three confidence-weighted ROVER variants using mean token log-probability, Shannon entropy, and Tsallis entropy; margin-weighted voting; and a learned per-word keep classifier trained via 5-fold group cross-validation (Supplementary Table S1). Fused WER used the same normalization pipeline.

Clinical Term Recall

To assess clinical relevance beyond aggregate WER, we measured the accuracy of medical terminology transcription using the Unified Medical Language System (UMLS). For each reference transcript, we identified all medical concept spans using QuickUMLS[17] (medspacy fork, threshold 0.8) against the UMLS 2025AB Metathesaurus, restricted to English and clinically relevant semantic types (drugs, conditions, symptoms, findings, procedures, anatomy, and lab results). QuickUMLS performs approximate string matching, returning single-word terms (e.g., “hypertension”) and multi-word phrases (e.g., “chest pain”) with their character positions. Common English words coincidentally matching UMLS abbreviations or eponyms were excluded via a curated stopword list (e.g., “said” matches Simian AIDS, “still” matches Still’s Disease). We aligned each model’s hypothesis to the reference using jiwer (rapidfuzz Levenshtein alignment) and identified which reference positions contained errors. A span was scored as an error if any of its words was incorrectly transcribed; clinical term recall is the proportion of spans correctly transcribed.

Cost Tracking

Cloud API costs were recorded from each provider’s billing dashboard at published rates; introductory credits covered the three lower-cost services, while the two most expensive, AWS and Google, exceeded their credits during the benchmark. On-device inference cost was considered zero per encounter. Per-encounter cost was total cost divided by encounter count; the annual illustration assumes one full-time clinician (~ 5,000 encounters per year).

AI Tool Use

This work used AI coding assistants (Anthropic Claude Code and Google Gemini Code Assist) for programming support in developing evaluation scripts, data-verification tools, and fusion code. All scientific design, dataset selection, analysis decisions, and manuscript text were authored and verified by the author; AI-generated code was reviewed and tested before use.

Results

On-Device vs. Cloud ASR Performance

On-device models approached or matched cloud API performance across all three datasets (Table 1, Fig. 1). We report standard and meaning-preserving WER side by side (see Methods). On the OSCE respiratory interview dataset, the best on-device model (whisper-distil-v3.5) trailed the best cloud API by 3.53 percentage points in standard WER (11.23% vs. 7.70%, Azure) but only 1.25 percentage points in meaning-preserving WER (8.51% vs. 7.26%, Deepgram). On PriMock57, the best on-device model (parakeet-tdt-0.6b-v2) trailed the best cloud API (AWS Transcribe Medical) by 5.17 percentage points in standard WER (13.85% vs. 8.68%) and 3.30 percentage points in meaning-preserving WER (11.39% vs. 8.09%). On the Kazi et al. psychiatric dataset, the gap was

small in both regimes: parakeet-tdt-0.6b-v2 (7.30% Std / 7.28% MP) trailed Azure (6.83% Std / 6.58% MP) by under one percentage point. Across the three datasets, the meaning-preserving-WER gap between the best on-device model and the best cloud API was 0.7 to 3.3 percentage points (vs. 0.5 to 5.2 in standard WER); the larger gap on conversational datasets reflects backchannel suppression, common in OSCE and primary-care recordings but rare in psychiatric interviews.

Table 1

Word error rate (%) by model and dataset, under two normalization regimes. Each dataset column has two sub-columns: **Std** uses the Whisper English text normalizer alone (the de facto leaderboard standard, used by the HuggingFace Open ASR Leaderboard and MLPerf), and **MP** is meaning-preserving WER, which layers on a normalization for clinical text (UK/US spelling, hyphenation, compound words, possessive eponyms, spaced acronyms, honorifics, dosing units, and conversational backchannels; full rules in Methods). Asterisked OSCE columns reflect the apostrophe-injected reference transcripts (see Methods). Parameter counts (millions) are approximate, drawn from model documentation and cross-checked against the deployed ONNX files; MedASR's count, which its distribution does not list, was measured directly from its ONNX model file. All on-device models were run as their published int8-quantized builds; in the Zipformer and Qwen3-ASR builds the upstream distribution leaves one component unquantized (the decoder and convolutional frontend, respectively). Cloud services do not disclose model sizes (Proprietary). Dataset sizes: OSCE n = 272, PriMock57 n = 57, psychiatric n = 71. ‡Google medical_conversation psychiatric WER is aggregated over 67 of 71 files; four of the six longest recordings had not completed within the 10-minute client-side wait and were not retried (see Methods).

Model	Type	Params (M)	OSCE Std*	OSCE MP*	PriMock57 Std	PriMock57 MP	Psychiatric Std	Psychiatric MP
Whisper distil-v3.5	On-device	756	11.23	8.51	14.19	11.71	7.35	7.33
Whisper turbo	On-device	809	12.01	9.37	15.53	13.05	7.55	7.54
Whisper base.en	On-device	74	14.27	11.79	20.27	18.08	7.64	7.63
Qwen3-ASR	On-device	600	12.37	9.68	14.74	12.33	7.94	7.92
Parakeet-TDT-0.6b-v2	On-device	600	17.06	14.56	13.85	11.39	7.30	7.28
SenseVoice (no ITN)	On-device	234	13.06	10.54	19.95	17.70	7.63	7.63
SenseVoice	On-device	234	12.80	10.26	21.01	18.72	7.68	7.69
Paraformer-en	On-device	220	17.95	16.77	29.12	27.33	9.28	9.20
NeMo FastConformer	On-device	114	22.73	20.11	32.34	30.00	11.28	10.92
Nemotron	On-device	600	35.59	33.76	38.16	36.49	37.40	37.45
Zipformer	On-device	67	33.90	33.27	36.61	35.57	11.14	10.84
MedASR	On-device	105	56.21	54.79	88.83	88.38	32.48	32.28
Azure Speech	Cloud (general)	Proprietary	7.70	7.34	9.88	9.35	6.83	6.58
Deepgram Nova-2 Medical	Cloud (medical)	Proprietary	10.10	7.26	11.80	9.14	6.99	6.89
AssemblyAI	Cloud (general)	Proprietary	10.29	7.60	13.76	11.26	7.81	7.81
AWS Transcribe Medical	Cloud (medical)	Proprietary	8.10	7.69	8.68	8.09	7.06	6.77
Google medical_conversation	Cloud (medical)	Proprietary	16.53	13.38	26.00	23.49	10.52‡	10.30‡

No single on-device model performed best across all datasets: whisper-distil-v3.5 led on OSCE, parakeet-tdt-0.6b-v2 on both PriMock57 and the psychiatric dataset. Among cloud APIs, Google medical_conversation showed substantially higher WER than the other four across all three datasets in both regimes (16.53% / 13.38% on OSCE, 26.00% / 23.49% on PriMock57, 10.52% /

10.30% on psychiatric; standard / meaning-preserving). Verbatim transcription of backchannels explains part of this gap: on OSCE, Google's output averaged 2,007 words per file against 1,314 in the reference (53% more tokens), with roughly 162 disfluency tokens per file versus 17 for Azure and 30 for AssemblyAI, consistent with the medical_conversation model being tuned to transcribe fillers and false starts verbatim rather than misconfigured. Google nonetheless remained the highest-WER cloud service after symmetric backchannel removal, indicating a residual content-word gap, not solely a disfluency artifact (see Limitations).

OSCE WER values use an apostrophe-injected reference (see Methods); this correction reduced every system's OSCE WER, by approximately 6 percentage points for most systems (1.6 to 6.9 overall), relative to the distributed reference.

ROVER Fusion

ROVER was evaluated on a pool of 9 competitive on-device models: all 36 pairs on PriMock57 and the psychiatric dataset, all 84 triples on PriMock57, and (see Methods) the targeted fusion of the two best solo models on OSCE. Fusion produced only small improvements over the best single on-device model (Fig. 2): the best pair gained 0.82 pp on PriMock57 (parakeet-tdt-0.6b-v2 + qwen3-asr, 13.03% vs. 13.85%), 0.22 pp on OSCE (parakeet-tdt-0.6b-v2 + sensevoice, 11.01% vs. 11.23% solo), and 0.15 pp on the psychiatric dataset. No fused result surpassed the best cloud API on any dataset (e.g., Azure 7.70% on OSCE).

On PriMock57, the only dataset with an exhaustive triple search, no triple beat the best pair: the best (13.12%) was 0.09 pp worse, with the third model introducing alignment noise that offset any voting benefit.

Clinical Term Recall

On-device models achieved clinical term recall (CTR) within 1 to 3 percentage points of cloud APIs across all datasets, with overlapping or near-adjacent 95% bootstrap CIs for the strongest comparisons (Table 2). On the psychiatric dataset, the best on-device models (whisper-turbo 92.0% [91.3, 92.6], whisper-distil 91.9% [91.2, 92.5], parakeet-tdt 91.8% [91.1, 92.4]) were within 1 percentage point of the best cloud API (Azure 93.0% [92.3, 93.5]). On the OSCE dataset, the best on-device model (whisper-turbo 93.6% [93.2, 93.9]) trailed the best cloud API (AWS Transcribe Medical 95.6% [95.4, 95.9]) by 2.0 percentage points, with non-overlapping intervals. On PriMock57, the gap was widest: parakeet-tdt (89.4% [88.3, 90.4]) trailed AWS Transcribe Medical (92.3% [91.5, 93.1]) by 2.9 percentage points. Across all cells, 95% CIs were 1 to 3 percentage points wide, meaning the best on-device-vs-cloud CTR differences are of similar magnitude to the measurement uncertainty.

Table 2

Clinical term recall (%) by model and dataset, with 95% confidence intervals. Intervals are bias-corrected and accelerated (BCa) bootstrap intervals over 10,000 resamples, using common random numbers within each dataset. Clinical term recall is the percentage of UMLS medical concept spans in the reference transcript transcribed without error. All five cloud services, including AWS Transcribe Medical, were evaluated on the complete datasets (Google on 396 of 400 files; see Methods).

Model	Type	OSCE (n = 272)	PriMock57 (n = 57)	Psychiatric (n = 71)
Whisper distil-v3.5	On-device	93.1 [92.8, 93.5]	88.5 [87.2, 89.5]	91.9 [91.2, 92.5]
Whisper turbo	On-device	93.6 [93.2, 93.9]	88.8 [87.7, 89.7]	92.0 [91.3, 92.6]
Whisper base.en	On-device	89.1 [88.6, 89.6]	80.3 [78.5, 81.8]	91.2 [90.6, 91.9]
Qwen3-ASR	On-device	91.9 [91.5, 92.4]	87.0 [86.0, 88.0]	90.9 [90.1, 91.6]
Parakeet-TDT-0.6b-v2	On-device	87.2 [86.3, 88.0]	89.4 [88.3, 90.4]	91.8 [91.1, 92.4]
SenseVoice (no ITN)	On-device	87.5 [86.9, 88.1]	78.1 [76.0, 79.6]	90.5 [89.8, 91.1]
SenseVoice	On-device	88.8 [88.3, 89.2]	77.8 [75.9, 79.3]	90.5 [89.8, 91.2]
Paraformer-en	On-device	76.9 [76.1, 77.7]	64.8 [62.5, 66.8]	88.1 [87.3, 88.8]
NeMo FastConformer	On-device	75.6 [74.7, 76.4]	63.8 [61.7, 65.8]	84.2 [83.4, 85.0]
Nemotron	On-device	60.0 [59.3, 60.8]	53.5 [52.0, 55.1]	52.4 [51.4, 53.3]
Zipformer	On-device	68.5 [67.7, 69.2]	57.7 [55.1, 60.1]	85.2 [84.3, 86.2]
MedASR	On-device	57.1 [53.7, 60.4]	14.4 [12.1, 17.0]	75.6 [71.6, 78.3]
Azure Speech	Cloud (general)	95.5 [95.2, 95.7]	89.8 [88.9, 90.7]	93.0 [92.3, 93.5]
Deepgram Nova-2 Medical	Cloud (medical)	95.4 [95.1, 95.6]	90.8 [90.0, 91.5]	92.4 [91.7, 93.0]
AssemblyAI	Cloud (general)	95.4 [95.2, 95.7]	91.4 [90.4, 92.3]	92.0 [90.9, 92.8]
AWS Transcribe Medical	Cloud (medical)	95.6 [95.4, 95.9]	92.3 [91.5, 93.1]	92.8 [92.2, 93.5]
Google medical_conversation	Cloud (medical)	91.0 [90.4, 91.5]	76.5 [74.3, 78.3]	89.0 [88.1, 89.8]

Model rankings differed between WER and CTR: whisper-turbo, second on overall WER, achieved the highest on-device CTR on two of three datasets. Medication names were the highest-error category for most models on every dataset, and the per-drug failures split sharply by deployment mode: pooled across the 12 on-device configurations, wellbutrin was misrecognized in 86% of occurrences (psychiatric), lisinopril in 92% and amoxicillin in 93% (PriMock57), and ramipril in 86%, rosuvastatin in 92%, and lisinopril in 85% (OSCE), versus 3% to 30% for the same drugs pooled across the five cloud services (per-model rates in Supplementary Table S4). The best overall on-device models were no exception: whisper-turbo erred on 50% to 78% of these occurrences, parakeet-tdt-0.6b-v2 on 83% to 100%. Drug names nonetheless remained the weakest cloud category on two of three datasets.

Cost Analysis

Total cloud API costs for the benchmark's approximately 90 hours of audio (400 encounters) ranged from \$19.33 (Azure Speech) to \$397.96 (AWS Transcribe Medical), with Deepgram Nova-2 Medical at \$25.78, AssemblyAI at \$28.10, and Google Cloud Speech at \$382.64 (Table 3). This corresponds to roughly \$0.05 to \$0.07 per encounter for Azure, Deepgram, and AssemblyAI versus about \$1.00 per encounter for Google and AWS, which are priced 14 to 20 times higher. For a single full-time clinician (about 5,000 encounters per year), annual costs at published rates would range from roughly \$240 to \$5,000 depending on the service. On-device inference incurred no per-encounter cost after initial model download (37 MB to 1.5 GB) and runs on consumer mobile devices clinicians already own.

Table 3
Cloud API cost for the full benchmark (~ 90 hours, 400 conversations). All five services were evaluated on the full benchmark (Google on 396 of 400 files; see Methods). On-device inference incurs no per-encounter cost.

Service	Model Type	Total Cost	Files Evaluated
Azure Speech	General	\$19.33	400
Deepgram Nova-2 Medical	Medical	\$25.78	400
AssemblyAI	General	\$28.10	400
AWS Transcribe Medical	Medical	\$397.96	400
Google Cloud Speech	Medical	\$382.64	396
On-device (any model)	General	\$0.00	400

Discussion

On-device ASR models reached clinical term recall within 1 to 3 percentage points of commercial cloud APIs across three datasets, making privacy-preserving on-device ASR a viable option for clinical documentation. The best-performing on-device models were general-purpose, with no medical fine-tuning, while three of the five cloud services deployed medical-specific models; a medical model adapted to conversational rather than dictation speech would likely narrow the gap.

No single on-device model dominated across all datasets and metrics: whisper-turbo achieved the highest clinical term recall on two of three datasets, and parakeet-tdt-0.6b-v2 the lowest WER on two of three. Model choice should depend on which metric matters and on deployment constraints, not WER alone.

Medication names were the highest-error category for most models, with the most severe per-drug failures concentrated on-device. MedASR suggests this on-device deficit is targetable: despite uncompetitive overall WER, this medical-dictation model erred on wellbutrin in only 5% of occurrences, versus 86% pooled on-device and near the 3% cloud rate (Supplementary Table S4). Prior work found that 5.7% of errors in speech-recognition-generated clinical documents are clinically significant, with medication the most common clinical error category.[18] This pattern mirrors the look-alike/sound-alike (LASA) medication-error phenomenon, which accounts for 6.2 to 14.7% of medication errors and is driven by measurable phonetic confusability.[19, 20] Mitigations include cross-referencing ASR output against patient medication lists and institutional formularies,[21] flagging phonetically similar drug names for mandatory clinician review,[19, 20] and requiring explicit confirmation before finalizing notes with high-risk medications.[18] Because drug names remained the weakest category for most cloud services as well, these safeguards apply to both on-device and cloud ASR.

ROVER fusion produced only small WER improvements over the best single on-device model (≤ 0.82 pp across datasets), and the best fused pair did not surpass the best cloud API on any dataset. None of six advanced fusion algorithms in the depth-study (Methods) exceeded naive equal-weight voting by more than ~ 0.4 percentage points, and the best practical method (epsilon-tuned CNC at 13.05% WER) remained nearly 4 percentage points above the oracle upper bound of 9.07%, the WER of an omniscient per-word selector (Supplementary Table S1). The residual gap reflects consensus errors across the available models rather than a limitation algorithmic fusion can recover.

For deployment, multi-model fusion roughly doubles compute and memory and adds latency without justifying the cost on-device; single-model improvement through medical fine-tuning, hotword biasing, or domain-specific pretraining is the more promising path.

On-device ASR eliminates patient audio transmission but not other privacy concerns: consent, transparency, data governance, and accountability still apply to both modes.[7, 8] In a 2025 study, patient willingness to consent to ambient documentation fell from 82% to 55% once AI use, data storage, and corporate involvement were disclosed, underscoring the need for transparent consent regardless of where audio is processed.[22] For privacy-sensitive specialties such as psychiatry, substance-abuse treatment, and reproductive health, eliminating audio transmission may be a prerequisite for patient acceptance and regulatory compliance

regardless of BAA protections;[7] clinical deployment may also fall under FDA oversight depending on intended use.[23] The zero-marginal-cost nature of on-device ASR has equity implications: a 2026 national study found ambient AI adoption in U.S. hospitals concentrated among institutions with stronger operating margins, larger size, and metropolitan location,[24] and annual cloud costs ranging from hundreds to several thousand dollars may reinforce this disparity. On-device models run on consumer hardware clinicians already own and may be the only legally compliant option where data-sovereignty laws are strict.

Implications for practice. On-device ASR is a viable option for ambient documentation, particularly in cost-, connectivity-, or sovereignty-constrained settings (community health centers, solo practices, jurisdictions with data localization requirements). Regardless of mode, off-the-shelf ASR is not yet trustworthy enough to elide a structured medication-reconciliation step before note finalization.

Future work should evaluate real clinical recordings such as the Bridge2AI-Voice corpus. OSCE examinations, held routinely worldwide and free of protected health information, are an under-used resource: institutional sharing under open licenses would substantially expand the available benchmark. ASR accuracy is only the first stage of the ambient-documentation pipeline: transcription errors propagate into the downstream LLM-generated note, where they may be amplified or masked. A recent Veterans Health Administration evaluation found AI-generated notes scored lower than human-produced notes across all quality domains,[25] and an interview study found clinicians edit AI drafts primarily for clinical accuracy, transcription errors, and missing context;[26] the clinical term recall gaps reported here therefore likely understate the downstream review burden. Specialty-specific benchmarks and pipeline-level evaluations of how transcription errors propagate through note generation would add more clinical signal than WER alone.

Several aspects merit caution. First and most important, all three datasets consist of simulated or enacted encounters rather than real patient-clinician conversations. Real encounters add overlapping speech, variable microphone placement, background noise, and non-lexical conversational sounds that degrade ASR accuracy, with reported word error rates reaching 50% in naturalistic multi-speaker settings;[6, 27, 28] the values reported here should therefore be read as a performance ceiling that would likely degrade in real-world deployment. No open-license datasets of real clinical conversations currently exist. The datasets also have narrow speaker diversity: the OSCE and PriMock57 recordings are UK-based and the psychiatric dataset uses student actors, so accents and dialects are limited, and on-device models may perform differently with more diverse or accented speakers.[9]

Other limitations are methodological. On-device models were evaluated using a single inference engine (sherpa-onnx); performance may vary under alternative runtimes. The UMLS-based term identification relies on a 36-token stopwords list curated by the author from high-frequency QuickUMLS collisions with common English words (e.g., “said”/Simian AIDS, “still”/Still’s Disease); as a single-reviewer artifact it may still miss collisions or exclude valid terms. Clinical term recall used span-level scoring, counting a span as an error if any of its words was misrecognized; because 78% of the 52,693 reference spans are single-word and the mean span length is 1.27 words (maximum five), this affects only a minority of multi-word phrases. WER is sensitive to backchannel handling and orthographic variants, which the meaning-preserving regime addresses (see Results). Finally, cloud results reflect early 2026 pricing and may change as providers update their rate schedules.

Conclusions

Privacy-preserving on-device ASR is a viable alternative to commercial cloud services for clinical conversation transcription, with clinical term recall within 1 to 3 percentage points of the best cloud service per dataset, standard WER within 0.5 to 5.2 percentage points, and meaning-preserving WER within 0.7 to 3.3 percentage points. The widest gaps occur on a single dataset (PriMock57); on the other two, meaning-preserving WER is within 1.3 percentage points of the best cloud service. Neither deployment mode reliably transcribes medication names, and on-device failures are far more frequent; until that improves, any clinical ASR deployment requires explicit medication verification before notes are finalized. On-device deployment lowers the cost barrier in under-resourced settings and keeps audio off the network in sensitive specialties; single-model medical fine-tuning is more promising than multi-model fusion as a path to closing the remaining gap with cloud services.

Abbreviations

ASR
automatic speech recognition
BAA
business associate agreement
BCa
bias-corrected and accelerated
CTR
clinical term recall
ITN
inverse text normalization
LASA
look-alike/sound-alike
MP
meaning-preserving
OSCE
objective structured clinical examination
PHI
protected health information
ROVER
recognizer output voting error reduction
UMLS
Unified Medical Language System
WER
word error rate.

Declarations

Ethics approval and consent to participate

This study did not constitute human subjects research and did not require institutional review board approval: it used only publicly available datasets of simulated or enacted clinical encounters, released by their original creators under open licenses, and did not involve human participants, identifiable private information, or protected health information.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are all publicly available under open licenses in their respective repositories. The OSCE respiratory interview dataset is available in the figshare repository, <https://doi.org/10.6084/m9.figshare.c.5545842.v1> (CC0 1.0).[29] PriMock57 is available in the Babylon Health GitHub repository, <https://github.com/babylonhealth/primock57> (CC-BY 4.0).[30] The psychiatric dataset is available in the Zenodo repository, <https://doi.org/10.5281/zenodo.4279041> (CC-BY 4.0).[13] The Unified Medical Language System (UMLS) 2025AB Metathesaurus is not redistributable and requires a free license from the U.S. National Library of Medicine.

Per-model inference outputs, aggregated WER and clinical term recall tables, and per-file fusion results for the exhaustive round-robin search and the fusion depth-study are deposited with the code repository described below.

All evaluation code, model inference scripts, the ROVER fusion implementation, the exhaustive pair and triple round-robin search, and the fusion depth-study (naive ROVER and six advanced fusion algorithms on PriMock57) are available under the MIT license. The current version is in the GitHub repository, <https://github.com/grai-lab/stapes>; the exact version reported here is archived in the Zenodo repository, <https://doi.org/10.5281/zenodo.22081372> (v1.0.0), with <https://doi.org/10.5281/zenodo.22081371> resolving to the latest archived version.[31] The depth-study scripts and per-file results are in the fusion_depth/ subdirectory. Cloud API evaluation scripts for each of the five services are included, along with the apostrophe-injection correction applied to the OSCE reference transcripts.

Competing interests

The author declares that they have no competing interests.

Funding

This study received no funding.

Authors' contributions

Following the CRediT (Contributor Roles Taxonomy): J.G.F.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing (original draft), Writing (review and editing), Visualization, Project administration. The author read and approved the final manuscript.

Acknowledgements

The author thanks the creators of the OSCE respiratory interview corpus, PriMock57, and the Kazi et al. psychiatric dataset for releasing these recordings under open licenses, which made this benchmark possible.

Use of artificial intelligence

This work used AI coding assistants (Anthropic Claude Code and Google Gemini Code Assist) for programming support in developing evaluation scripts, data-verification tools, and fusion code, as described in Methods. All scientific design, dataset selection, analysis decisions, and manuscript text were authored and verified by the author; AI-generated code was reviewed and tested before use.

References

1. Sahni NR, Carrus B. Artificial intelligence in U.S. health care delivery. *N Engl J Med*. 2023;389(4):348–58. 10.1056/NEJMr2204673.
2. Shah SJ, Devon-Sand A, Ma SP, Jeong Y, Crowell T, Smith M, et al. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. *J Am Med Inf Assoc*. 2025;32(2):375–80. 10.1093/jamia/ocae295.
3. Stults CD, Deng S, Martinez MC, Wilcox J, Szwedinski N, Chen KH, et al. Evaluation of an ambient artificial intelligence documentation platform for clinicians. *JAMA Netw Open*. 2025;8(5):e258614. 10.1001/jamanetworkopen.2025.8614.
4. Olson KD, Meeker D, Troup M, Barker TD, Nguyen VH, Manders JB, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Netw Open*. 2025;8(10):e2534976. 10.1001/jamanetworkopen.2025.34976.
5. Haberle T, Cleveland C, Snow GL, Barber C, Stookey N, Thornock C, et al. The impact of Nuance DAX ambient listening AI documentation: a cohort study. *J Am Med Inf Assoc*. 2024;31(4):975–9. 10.1093/jamia/ocae022.

6. 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 Med Inf Decis Mak.* 2025;25(1):236. 10.1186/s12911-025-03061-0.
7. Elsayed N. Socio-technical risks of clinical speech-to-text systems: transparency, privacy, and reliability challenges in AI-driven documentation. *Int J Med Inf.* 2026;214:106419. 10.1016/j.ijmedinf.2026.106419.
8. Anderson TN, Mohan V, Gold JA. Ethical considerations for clinical adoption of ambient digital scribe technology. *J Am Med Inf Assoc.* 2026;33(3):770–5. 10.1093/jamia/ocaf227.
9. Afonja T, Olatunji T, Ogun S, Etori NA, Owodunni A, Yekini M. Performant ASR models for medical entities in accented speech. In: *Proceedings of Interspeech 2024.* Kos: ISCA; 2024. pp. 2315-9. 10.21437/Interspeech.2024-2261
10. Nolan EJ, Burke HB. Accuracy of large language model transcription of simulated physician-patient verbal interactions. *BMC Med Inf Decis Mak.* 2026;26(1):124. 10.1186/s12911-026-03414-3.
11. Fareez F, Parikh T, Wavell C, Shahab S, Chevalier M, Good S, et al. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Sci Data.* 2022;9(1):313. 10.1038/s41597-022-01423-1.
12. Papadopoulos Korfiatis A, Moramarco F, Sarac R, Savkov A. PriMock57: a dataset of primary care mock consultations. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (volume 2: short papers).* Dublin: Association for Computational Linguistics; 2022. pp. 588 – 98. 10.18653/v1/2022.acl-short.65
13. Kazi N, Kuntz M, Kanewala U, Kahanda I. Dataset for automated medical transcription. Zenodo. 2020. <https://doi.org/10.5281/zenodo.4279041>
14. Srivastav V, Zheng S, Bezzam E, Le Bihan E, Koluguri NR, Želasko P, et al. Open ASR leaderboard: towards reproducible and transparent multilingual and long-form speech recognition evaluation. *arXiv [Preprint].* 2025. 10.48550/arXiv.2510.06961.
15. Fiscus JG. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In: *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding.* Santa Barbara: IEEE; 1997. pp. 347 – 54. 10.1109/ASRU.1997.659110
16. Mangu L, Brill E, Stolcke A. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Comput Speech Lang.* 2000;14(4):373–400. 10.1006/csla.2000.0152.
17. Soldaini L, Goharian N. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In: *Proceedings of the Medical Information Retrieval Workshop (MedIR) at SIGIR 2016.* Pisa; 2016.
18. 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 Netw Open.* 2018;1(3):e180530. 10.1001/jamanetworkopen.2018.0530.
19. Lambert BL, Dickey LW, Fisher WM, Gibbons RD, Lin SJ, Luce PA, et al. Listen carefully: the risk of error in spoken medication orders. *Soc Sci Med.* 2010;70(10):1599–608. 10.1016/j.socscimed.2010.01.042.
20. Bryan R, Aronson JK, Williams A, Jordan S. The problem of look-alike, sound-alike name errors: drivers and solutions. *Br J Clin Pharmacol.* 2021;87(2):386–94. 10.1111/bcp.14285.
21. Rash-Foanio C, Galanter W, Bryson M, Falck S, Liu KL, Schiff GD, et al. Automated detection of look-alike/sound-alike medication errors. *Am J Health Syst Pharm.* 2017;74(7):521–7. 10.2146/ajhp150690.
22. Lawrence K, Kuram VS, Levine DL, Sharif S, Polet C, Malhotra K, et al. Informed consent for ambient documentation using generative AI in ambulatory care. *JAMA Netw Open.* 2025;8(7):e2522400. 10.1001/jamanetworkopen.2025.22400.
23. Warraich HJ, Tazbaz T, Califf RM. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *JAMA.* 2025;333(3):241–7. 10.1001/jama.2024.21451.
24. Yang F, Graetz I. Ambient AI tool adoption in US hospitals and associated factors. *Am J Manag Care.* 2026;32(1):e25–30. 10.37765/ajmc.2026.89876.
25. Reddy A, Gunnink E, Wheat CL, Pawlikowski S, Payne CM, Wiltz S, et al. Rapid evaluation of artificial intelligence technology used for ambient dictation in primary care: comparing the quality of documentation of artificial intelligence-generated and human-produced clinical notes. *Ann Intern Med.* 2026;179(6):765–72. 10.7326/ANNALS-25-02772.

26. Guo Y, Hu D, Yang Z, Chow E, Tam S, Perret D, et al. Clinicians’ rationale for editing ambient AI-drafted clinical notes: persistent challenges and implications for improvement. *J Am Med Inf Assoc.* 2026;33(7):1345–53. 10.1093/jamia/ocag059.

27. Tran BD, Latif K, Reynolds TL, Park J, Elston Lafata J, Tai-Seale M, et al. Mm-hm, Uh-uh: are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology? *J Am Med Inf Assoc.* 2023;30(4):703–11. 10.1093/jamia/ocad001.

28. Alboksmaty A, Aldakhil R, Hayhoe BW, Ashrafian H, Darzi A, Neves AL. The impact of using AI-powered voice-to-text technology for clinical documentation on quality of care in primary care and outpatient settings: a systematic review. *EBioMedicine.* 2025;118:105861. 10.1016/j.ebiom.2025.105861.

29. Fareez F, Parikh T, Wavell C, Shahab S, Chevalier M, Good S et al. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *figshare.* 2022. <https://doi.org/10.6084/m9.figshare.c.5545842.v1>

30. Papadopoulos Korfiatis A, Moramarco F, Sarac R, Savkov A. PriMock57: primary care mock consultation dataset. *GitHub.* 2022. <https://github.com/babylonhealth/primock57>. Accessed 24 Aug 2026.

31. Faulkenberry JG. stapes: an open benchmark of on-device and cloud speech recognition for clinical conversations (v1.0.0). *Zenodo.* 2026. <https://doi.org/10.5281/zenodo.22081372>

Figures

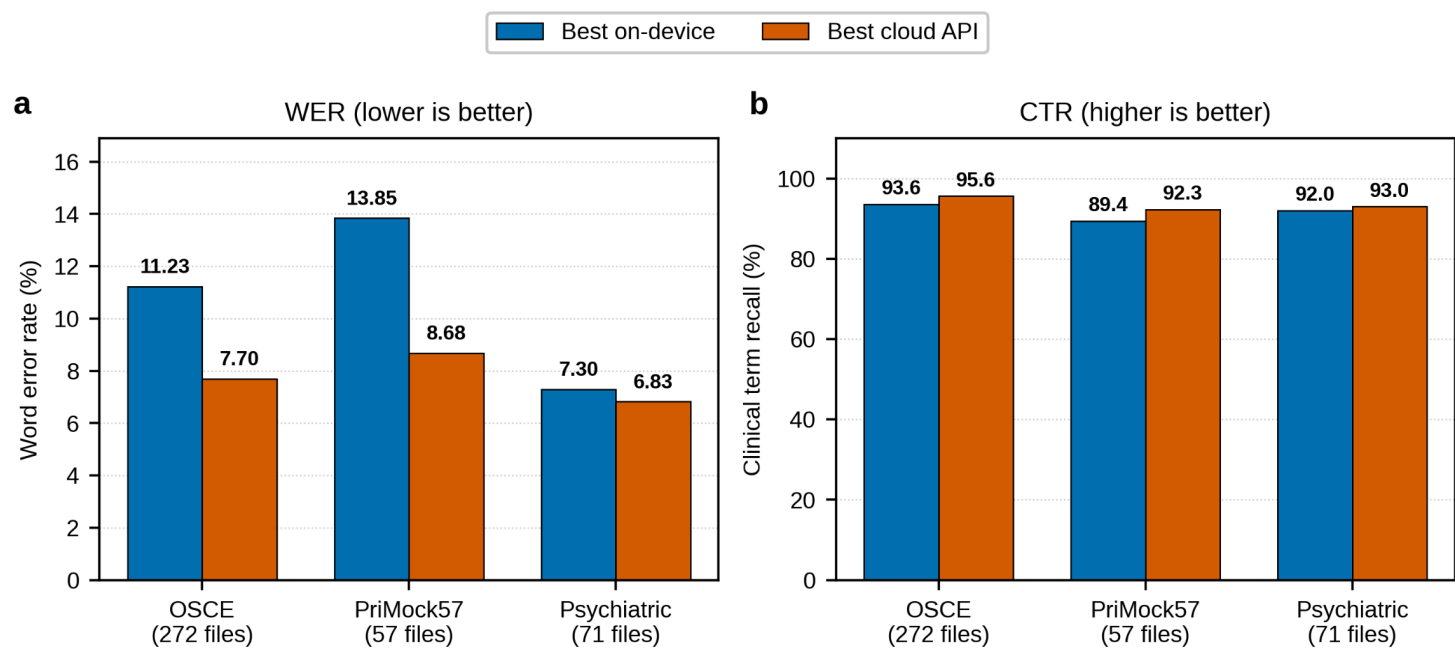


Figure 1

Best on-device model versus best cloud API, per dataset. Word error rate and clinical term recall of the best on-device ASR model against the best cloud API on each dataset. **a**, Word error rate (%; standard Whisper-normalized regime; see Methods) of the best on-device model and the best cloud API on each of the three clinical conversation datasets (OSCE respiratory interviews, n = 272; PriMock57 primary care, n = 57; Kazi et al. psychiatric, n = 71). Lower is better. **b**, Clinical term recall (%) of the best on-device model and the best cloud API on each dataset. Higher is better. Clinical term recall was computed as the proportion of UMLS medical concept spans in the reference transcript that were correctly transcribed.

Alt text: Two-panel grouped bar chart comparing the best on-device ASR model with the best cloud API on each of three clinical conversation datasets. Panel a shows word error rate (lower is better); panel b shows clinical term recall (higher is better). Datasets are OSCE respiratory interviews (n=272), PriMock57 primary care (n=57), and Kazi et al. psychiatric (n=71).

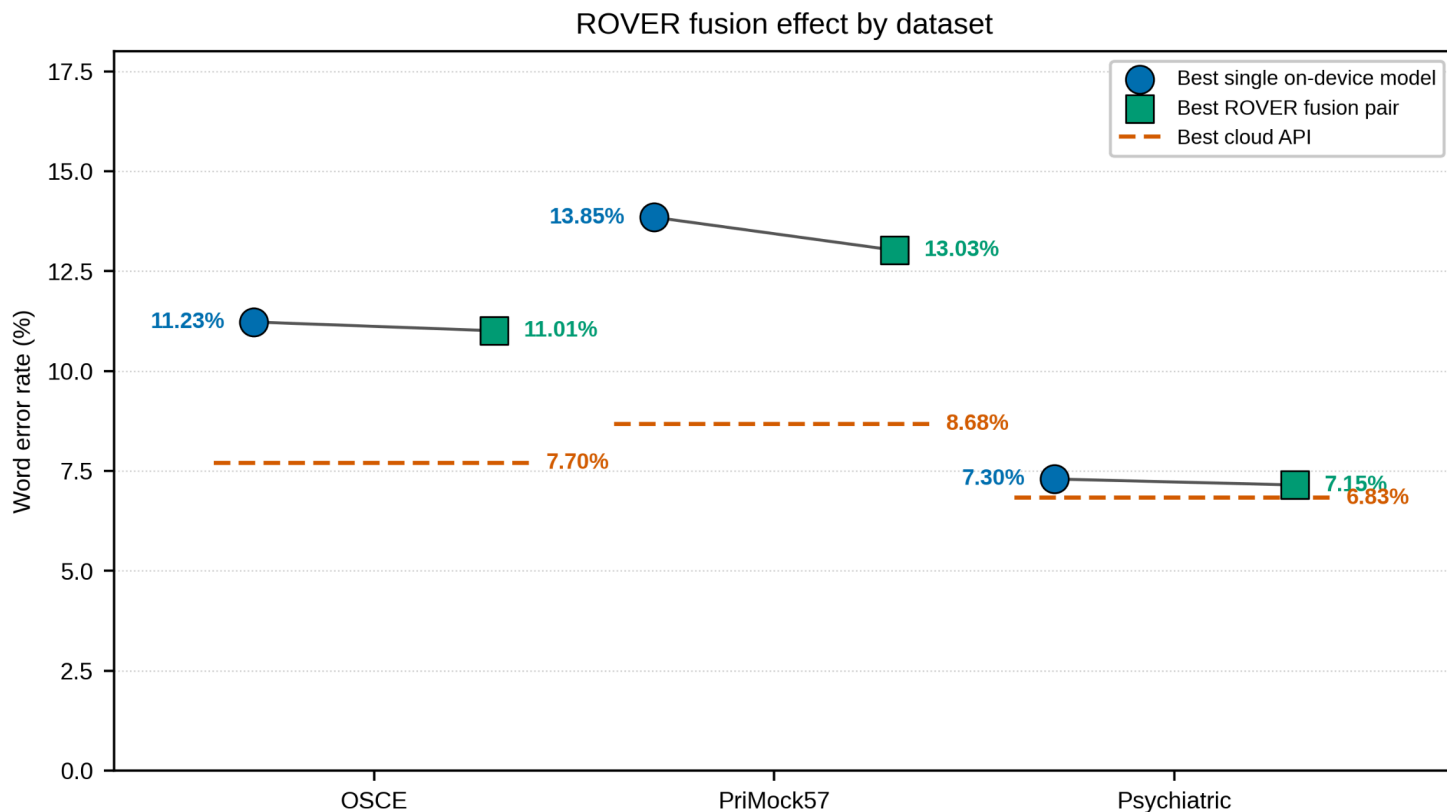


Figure 2

ROVER fusion yields only small gains over a single on-device model. The best cloud API is drawn as a reference. For each dataset, the figure shows the word error rate of the best single on-device model (left point), the best two-model ROVER fusion pair (right point, connected by a grey line to display the fusion delta), and the best cloud API as a dashed horizontal reference line. Fusion provided ≤ 0.82 percentage point improvements on all three datasets. On the OSCE respiratory interview dataset, the best fused on-device pair (parakeet-tdt-0.6b-v2 + sensevoice, 11.01%) did not surpass the best cloud API (Azure, 7.70%).

Alt text: Dot-and-line plot showing word error rate for the best single on-device model versus the best two-model ROVER fusion pair on each of three clinical conversation datasets, with the best cloud API drawn as a dashed horizontal reference line. Fusion produces small improvements over the best single on-device model on all three datasets but does not reach the best cloud API on any of them.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [stapessupplementary.docx](#)