Second Thoughts: Revisable Trailing Context in LLM-Grounded Clinical Speech

Dave Makhervaks
Ambience Healthcare

Vinesh Gudla
Ambience Healthcare

Kabir Mahal
Ambience Healthcare

Aditya Prabhakar
Ambience Healthcare

Josh Cowdy
Ambience Healthcare

Eric Hunter
Ambience Healthcare

Raizy Leizerowski
Ambience Healthcare

Ryan Schwers
Ambience Healthcare

Madeline Grade
Ambience Healthcare

Tejaswi Tenneti
Ambience Healthcare

Abstract

Clinical dictation is the upstream signal for an expanding class of LLM-driven healthcare systems namely note-generation pipelines, clinical-decision-support agents, structured-data extractors. Errors in the transcript silently propagate through every downstream LLM call. We present a production pipeline whose main contribution is the combination of multi-source ASR ensembling with LLM reconciliation, in a rolling-window architecture. This gives the reconciler revisable trailing context where a frontier LLM fuses parallel transcripts from heterogeneous fine-tuned ASR backends grounded in structured patient context (medications, problems, clinician identity) through a sliding K-buffer audio re-pass. This lets it correct prior output as new audio arrives. Unlike other multi-ASR fusion work, no word-level voting or confusion-network construction is used in our approach. A prompt transfer ablation shows that the gain comes from the audio re-pass architecture improvements, rather than from how the reconciler is prompted. On a 226-encounter clinical dictation dataset, ensembling-plus-reconciliation is the dominant contribution: it reduces normalized keyword error by 40% relative to a single fine-tuned backend with post-processing, with 95% CIs. The rolling-window variant adds a further improvement across raw and normalized metrics (16–26% relative vs. a non-revisable mini-batch ensemble; all $p < 10^{-3}$ under encounter-level bootstrap). In deployed clinician use, medication-name correction edit rate dropped 51% (12.21% → 6.04%) post-rollout. We additionally characterize what the LLM-in-the-loop reconciler is sensitive to (structured grounding, multi-source diversity) and what it is not (prompt engineering alone, single-backend configurations). We discuss implications for the broader design of LLM systems that operate on noisy upstream signals.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD'26 Workshop on SciSoc Agents & LLMs, Jeju, South Korea
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

CCS Concepts

• **Computing methodologies** → **Speech recognition**; *Natural language processing*; Artificial intelligence.

Keywords

clinical NLP, medical ASR, grounded LLMs, ensemble methods

ACM Reference Format:
Dave Makhervaks, Vinesh Gudla, Kabir Mahal, Aditya Prabhakar, Josh Cowdy, Eric Hunter, Raizy Leizerowski, Ryan Schwers, Madeline Grade, and Tejaswi Tenneti. 2026. Second Thoughts: Revisable Trailing Context in LLM-Grounded Clinical Speech. In *Proceedings of KDD'26 Workshop on SciSoc Agents & LLMs: Agentic AI for Scientific and Societal Advances (KDD'26 Workshop on SciSoc Agents & LLMs).* ACM, New York, NY, USA, 11 pages.

1 Introduction

Clinical documentation is one of the largest sources of administrative burden in medicine: physicians spend 35–50% of the workday on documentation tasks [1]. Speech-driven workflows such as dictation and ambient capture are increasingly the entry point for LLM-driven health systems: note generators, clinical-decision-support agents, structured-data extractors and summary generators. Dictation-based note entry alone materially improves note timeliness in residency settings [2]. The broader voice-AI healthcare market is projected to grow substantially over the next decade [3].

A consequence rarely emphasized is that every downstream LLM system is conditioned on a noisy speech-to-text artifact whose errors are not visible to the next stage. A misheard medication name flows verbatim into the generated note. A garbled dosage propagates into structured order entry. A misrecognized abbreviation triggers an incorrect clinical-decision-support alert. The downstream LLM cannot recover the information loss incurred by the STT model. Thus, the quality of the upstream transcript is critical for the entire healthcare-LLM stack and improving it has compounding effects on every system built on top. Recent work measures this directly where medical-dialogue summarization quality degrades substantially under realistic ASR error rates [19].

We present a production clinical-dictation pipeline that treats the transcript as revisable evidence rather than append-only text. The pipeline combines three design choices. First, it uses a domain-adapted encoder-decoder ASR model fine-tuned on clinical dictation, because generic ASR systems remain brittle on long-tail

clinical terminology, medications, and specialty-specific phrasing. Second, it runs an ensemble of ASR backends in parallel and uses a frontier LLM to reconcile their raw transcripts using structured patient context, including medications, problems, clinician identity, and note context. Third, the deployed system uses a rolling-window architecture in which each tick reprocesses a sliding K-buffer of recent audio, allowing the reconciler to revise recent output when later audio resolves an earlier ambiguity.

The central design claim is that, in LLM-in-the-loop systems built on noisy upstream signals, when prior output remains revisable matters as much as how the LLM is prompted. A mini-batch pipeline can pass previous text as context, but once a chunk's audio has left scope, later calls cannot recover from an early chunk-boundary error. A rolling audio re-pass changes the evidence available to the reconciler: overlap regions can be re-transcribed and corrected as new context arrives. We isolate this architectural property with a prompt-transfer ablation, showing that porting rolling-window prompt features into a non-revisable mini-batch architecture does not close the gap. This paper makes four contributions:

(1) A production multi-source clinical ASR + LLM reconciliation architecture. We combine a domain-adapted clinical ASR backend with a medical-specialized backend and use an LLM reconciler grounded in structured patient context to fuse raw parallel transcripts without word-level voting or confusion-network construction.
(2) An empirical evaluation of the pipeline design space. On a 226-encounter clinical dictation evaluation spanning 28 specialties and 15 clinicians, we compare streaming ASR, single-backend LLM post-processing, full-batch multi-source reconciliation, mini-batch reconciliation, and rolling-window reconciliation. Multi-source reconciliation is the dominant source of improvement, reducing normalized keyword error by 40% relative to a single fine-tuned backend with post-processing.
(3) A revisable-context ablation separating architecture from prompting. We evaluate a rolling K-buffer audio re-pass that lets the LLM revise prior output with replace-not-append semantics. Compared with a non-revisable mini-batch ensemble using the same backends, reconciler, and grounding context, the rolling-window variant improves raw WER by 22.3% and raw KER by 16.3%. A prompt-transfer ablation shows that the gain comes from keeping the relevant audio in scope, not from prompt wording alone.
(4) A production validation and sensitivity characterization. We characterize the system's sensitivity to structured grounding and backend diversity, and we report deployed edit telemetry across 218 clinicians and 20 specialties. After rollout, medication-name correction edits dropped 51% relative, from 12.21% to 6.04%, with the largest improvements occurring in categories targeted by the multi-source, clinically grounded architecture.

Together, these results support a broader systems lesson: for healthcare LLM pipelines that operate on noisy upstream signals, transcript generation should not be treated as a one-shot preprocessing step. The scope of evidence that remains revisable across calls is a first-class architectural choice, with direct consequences for downstream LLM reliability.

2 Related Work

Medical ASR.. Large-scale general-purpose ASR systems have established strong general transcription baselines [4], and commercial speech-to-text APIs make such capabilities broadly available [5]. However, clinical dictation stresses a different error profile: long-tail medication names, abbreviations, procedures, proper nouns, and specialty-specific phrasing [10]. Medical-specialized ASR systems address part of this gap by training on clinical speech, but can remain uneven across entity categories, especially for names and other entities absent from de-identified training data [7]. Commercial clinical recognizers provide mature dictation products, but are typically delivered as closed products or APIs rather than open, fine-tunable checkpoints for custom LLM reconciliation pipelines [8, 9].

Streaming vs. batch ASR.. The streaming-vs-batch distinction is well-studied in general settings [11], but its downstream implications for LLM-in-the-loop pipelines have received little attention. Streaming models operate causally, emitting tokens before observing future audio context. This can be especially harmful for medical terms with phonetic prefixes (e.g., "hyper-" vs. "hypo-").

LLM-based reconciliation and post-processing. HyPoradise [13] processes N-best hypotheses from a single ASR backend at a time via an LLM; it has since anchored an active line of post-ASR LLM-correction work, including the GenSEC community challenge [20] and recent SOTA NeKo [21], all of which operate on N-best lists from one backend rather than parallel transcripts from multiple. Similar to our approach, Prakash et al. [17] fuse multiple ASR backends through an LLM rather than purely by voting. In their design, the hypotheses are first aligned into a word-level confusion network, and an LLM is invoked downstream. Three concrete properties distinguish our system from theirs: (a) the LLM performs alignment and arbitration on raw parallel hypotheses directly (no alignment or confusion-network) (b) reconciliation is chunk-incremental and grounded in structured patient context (c) the trailing context is revisable via a rolling audio re-pass, letting the LLM correct prior output as new audio resolves earlier ambiguity rather than committing decisions on a full-audio batch. Generative error correction has also been shown to reduce clinical transcription errors when coupled to dictation systems [13, 14], and earlier ROVER-style fusion [12] performs voting purely at the word level without LLM grounding.

Rolling re-transcription. Macháček et al.'s LocalAgreement-2 [15] explore repeatedly re-decoding a rolling audio buffer with a single ASR backend and commits only the longest common prefix shared by consecutive hypotheses. Simul-Whisper [18] pursues the same single-backend streaming objective with an attention-guided truncation and commit policy. Both systems make algorithmic commit decisions over a single ASR hypothesis stream. Our architecture generalizes the rolling re-transcription pattern to multi-backend parallel dispatch with LLM-based content alignment, replacing algorithmic prefix commitment with a semantically revisable trailing window. This design is appropriate when a clinically grounded

Table 1: The four operating modes. "Incr." = incremental output. "Rev." = revisable trailing context (the reconciler can correct prior output as new audio arrives).

Mode	Incr.?	Rev.?	First-emit lat.
Full-batch	No	N/A	Highest
Rolling-window	Yes	Yes	Med.
Mini-batch	Yes	No	Low
Pure streaming	Yes	No	Lowest

LLM can semantically align overlapping hypotheses and arbitrate among heterogeneous sources.

Voice Activity Detection for long-form ASR.. Bain et al.'s pipeline [16] performs two-stage Voice Activity Detection (VAD) smoothing tuned to feed a single fixed-window backend's training-audio length. We use VAD differently, as a chunker whose knobs govern how much audio each reconciler call sees.

3 System

The pipeline supports four operating modes spanning a quality/latency Pareto frontier (Table 1).

3.1 Multi-Source ASR Input

The pipeline runs multiple ASR backends in parallel (fusion at every call, not dynamic tool selection) and supplies their outputs to the reconciler LLM. Our main configurations use two; we evaluate a three-backend variant in §6.3. We refer to backends by role:

Backend FT is a large-scale multilingual encoder-decoder Transformer ASR pre-trained on diverse weakly supervised audio, fine-tuned on clinical dictation, and self-hosted behind a streaming-capable inference service. It supports word-level timestamps and is the primary domain-adapted recognizer. Fine-tuning is a prerequisite: without it, downstream LLM reasoning degrades because the input signal is too noisy on medical vocabulary.

Backend Med is a ∼100M-parameter Conformer-based medical-specialized ASR trained on de-identified clinical dictation. Although it is lightweight, it excels on medical terminology and medication names. However, it underperforms on proper names due to de-identified training. The reconciler treats it as a strong medical-vocabulary signal whose proper-name output should be discounted.

The backends share a common interface (raw PCM in, plain-text transcript out), so any architecture can substitute one for another without code changes. We exploit this for the backend-composition ablation (§6.3).

3.2 LLM Reconciliation with Clinical Grounding

The reconciler is a single LLM call per chunk (mini-batch) or per tick (rolling-window). It receives: (1) the current segment's transcripts from both backends (2) the previously-reconciled trailing text (a rolling textual context chain) and (3) a structured patient/encounter context block including medication list, problem list, and clinician identity as a read-only block. The call produces a single corrected transcript covering the current segment.

A single prompt governs the reconciliation step. The system prompt carries three classes of instructions: formatting normalization (units, numerals, spoken punctuation, capitalization), reconciliation rules (which backend to prefer for which content type, how to treat overlap), and boundary handling (preserving the insertion-point capitalization and punctuation given the surrounding text). The user provides the inputs: the newly dictated audio, the trailing context, the identity block, and the text immediately before and after the cursor. A small deterministic post-step trims any echoed cursor context and reattaches a frozen prefix when a sliding text window is in use.

The rolling textual context means each call receives the prior segment's reconciled output as part of its input, enabling narrative coherence across segments. The identity-context block is the single largest non-architectural contributor to quality. Disabling it regresses the names category by roughly 5× (∼11% → ∼50% keyword error rate - KER - on names), driven almost entirely by misspelled patient and provider names. We leave it on in every configuration evaluated below. This is consistent with a broader observation: LLM reconciliation works best when the LLM has *a grounding signal that neither input source provides*. In our case, the clinical context lets the LLM pick the better of two transcript hypotheses by reference to the patient's context.

3.3 Mini-Batch Reconciliation (Non-Revisable Baseline)

The non-revisable baseline (Fig. 1, left): audio is split into VAD-aligned chunks, each transcribed by two backends in parallel, and reconciled by an LLM. Chunks process sequentially; each reconciliation receives a corrected text trailing window (up to 250 tokens) from prior chunks. The reconciler has limited, non-revisable upstream content —the trailing context carries only *text*, not *audio*.

Latency. Each tick is $\max(t_A, t_M) + t_{\text{LLM}}$ (ASRs in parallel, reconciler sequential). Median per-tick is ∼1.8 s in our deployment.

Limitation that motivates the next architecture. Once a chunk is reconciled, its audio is out of scope. If a chunk-boundary mishears a clinical term, no later step can recover it.

3.4 Rolling-Window Reconciliation (Revisable Trailing Context)

The revisable variant of the reconciliation architecture (Fig. 1, right). Audio is split into VAD-aligned buffers via the same VAD-based chunking front-end, but instead of feeding ONE buffer per ASR call the wrapper sends a **rolling K-buffer audio window** (we use K=3). Each tick re-transcribes the trailing portion of the prior reconciled text plus whatever new audio has been captured since. The reconciler identifies the overlap by content alignment and produces a single corrected output covering both the revised overlap region and the new content. Macháček et al.'s LocalAgreement-2 [15] explored rolling audio re-decodes from a single ASR backend, committing only the longest common prefix shared by consecutive hypotheses. We generalize the rolling re-transcription pattern to multi-backend parallel dispatch with LLM-based content alignment, replacing algorithmic prefix commitment with replace-not-append semantics over a revisable trailing window.

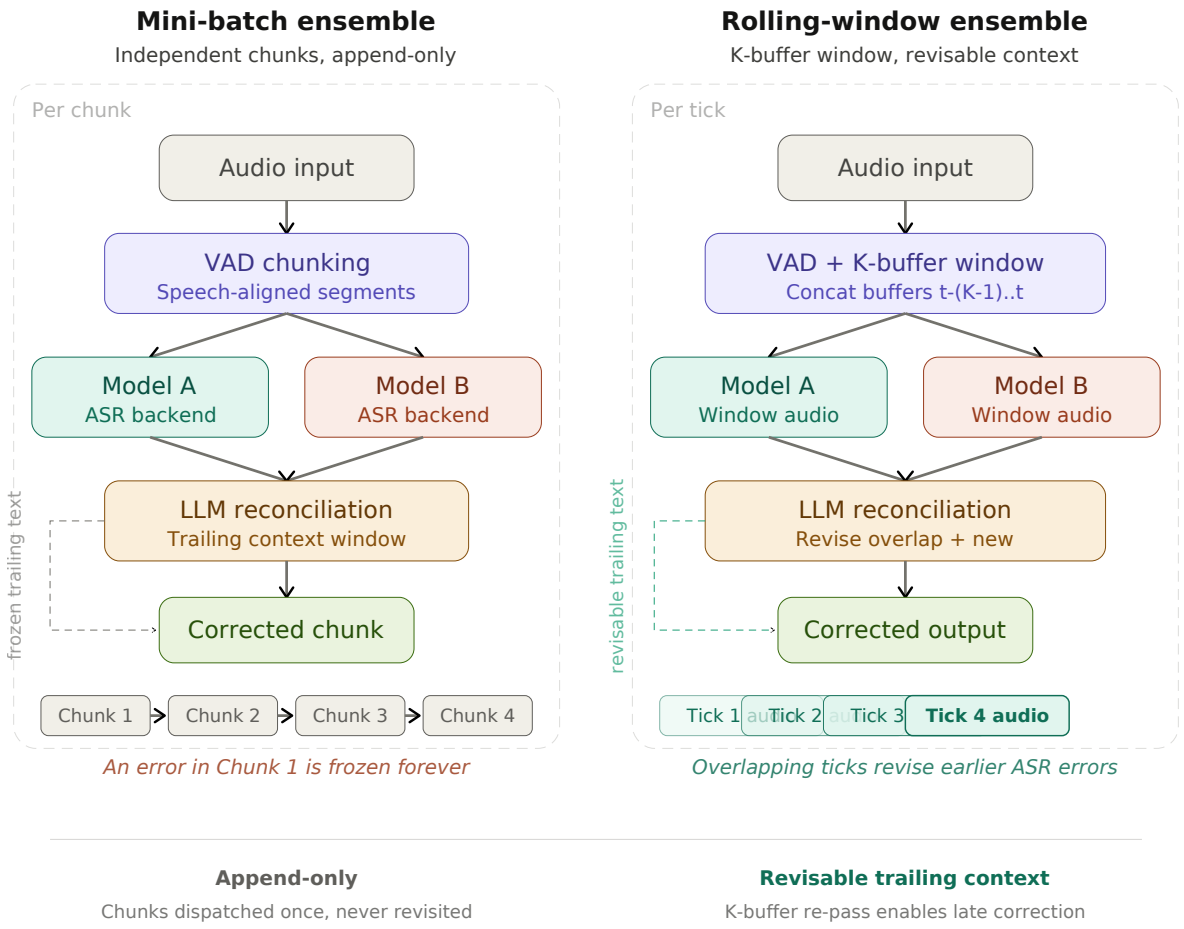

Figure 1: *Left:* mini-batch dispatches each VAD-aligned chunk independently and reconciles with a *frozen text trailing window*—chunk-boundary errors are not recoverable. *Right:* rolling-window dispatches a sliding K-buffer audio window per tick so the reconciler sees overlapping audio across ticks and can revise prior output as more context arrives.

Two structural defenses against bulk duplication. Naive prompts can cause the LLM to bulk-rewrite "previously processed" text on every tick, worsening performance and risking silent duplication:

- **Hard `max_tokens` cap per call**, derived from the total input length plus a small buffer. The model cannot emit more than its input plus that buffer, making bulk re-rendering of trailing context structurally impossible.
- **Adaptive K shrinking**: the wrapper drops the oldest buffer when the concatenated audio exceeds Backend FT's per-call limit. The newest buffer is always preserved.

Input-field segregation. Prior reconciled output enters the prompt as a *separate read-only context field*. Earlier iterations of this prompt design confirmed this is a structural decision. In past experiments when reconciled context replaced the primary transcript fields, the normalized KER tripled and worst-case encounters degraded from 0% to 58% KER.

Prompt design. A single coherent objective of producing the corrected full transcript covering both the revised previously-processed content and the new content governs the reconciliation. Further, a clinical-context inference clause licenses the LLM to repair garbled tokens that do not match any recognizable concept using surrounding clinical context (e.g., recovering "TMJ" when both backends mishear it as a non-existent abbreviation), without licensing aggressive rewrites of well-recognized content.

Latency. Each tick is $\max(t_A, t_M) + t_{\text{LLM}}$, but t_{LLM} dominates because the prompt carries the visible window plus K buffers' worth of transcripts. Median per-tick is ~2.5 s in our deployment—~0.7 s slower than mini-batch. The ASR latency is essentially unchanged.

4 Experimental Setup

Dataset. We evaluate on a 226-encounter clinical dictation dataset spanning 28 medical specialties and 15 clinicians, each able to dictate across multiple specialties. Encounters were recorded

in varied settings (home, hospital, office) by English speakers spanning a range of accents. Durations range from ∼30 s to ∼8 min. Reference transcripts are produced by multiple expert medical annotators in multi-pass review against a strict taxonomy.

Metrics. We report Word Error Rate (WER) and a domain Keyword Error Rate (KER), each computed per encounter i.
WER is the standard word edit-distance rate:

$$\mathrm{WER}_i \;=\; \frac{S_i + D_i + I_i}{N_i},$$

where S_i, D_i, I_i are substitutions, deletions, and insertions from a Levenshtein word alignment and N_i is the reference word count.
KER is the fraction of reference keyword occurrences missing from the hypothesis. For a unique keyword k, with non-overlapping occurrence counts $c^{\mathrm{ref}}_{i,k}$ and $c^{\mathrm{hyp}}_{i,k}$ in the reference and hypothesis,

$$\mathrm{KER}_i \;=\; \frac{\sum_k \max\!\left(0,\; c^{\mathrm{ref}}_{i,k} - c^{\mathrm{hyp}}_{i,k}\right)}{\sum_k c^{\mathrm{ref}}_{i,k}}.$$

We bucket words into their respective categories (medications, medical terms, proper names, dates, numerical values, devices) using an LLM-based structured entity extractor.

Both metrics are reported in raw (case and punctuation-sensitive) and normalized (lowercased, punctuation stripped, whitespace collapsed) form. All aggregates are micro across the corpus. Unless stated, results are means of three trials per configuration; noise floor is ≤ 0.04 pp on Norm. KER and $\sim \pm 0.01$–0.02 on WER at reconciler temperature 0.

KER extractor validity. Many of our clinical claims are based on KER, thus the extractor's accuracy warrants explicit discussion. The category taxonomy was defined in collaboration with medical experts, and extractor outputs were aligned against expert-annotated reference transcripts by clinically-trained reviewers. Since the same entity extractor tags both sides of every KER comparison, systematic extractor recall/precision errors are correlated across the two sides and largely cancel in the per-encounter difference. As an independent, extractor-free cross-check, WER tracks KER directionally and proportionally across every main comparison e.g., S → S+PP: WER −45% alongside KER −57%; MBE → RWE: WER −22.3% alongside KER −16.3%. KER gains driven predominantly by extractor artifacts would not produce co-moving WER at this magnitude.

Direct audit against human labels. We additionally validate the extractor directly against human-labeled entity spans on a 122-encounter audit. We pool across all categories, the extractor achieves an F1 of 0.865 under substring matching. We choose substring matching, since many of the strict-match disagreements are boundary/granularity differences (e.g., *Plummer* vs *Ms. Plummer*, *carboplatin* vs *induction carboplatin*). This is within the F1 range Durango et al. [24] report for high-performing clinical NER systems on private corpora (0.75–0.96, Table 2).

Configurations. We evaluate six configurations:

- **S**: streaming Backend FT with no post-processing.
- **S+PP**: streaming Backend FT with the LLM reconciler applied as post-processing.
- **B+PP**: batch Backend FT with the same LLM reconciler as post-processing. The single-backend ceiling.

Table 2: Raw single-model ASR on the 226-encounter eval. No post-processing or ensembling. Lower is better.

Model	WER	KER	Norm. WER	Norm. KER
GPT-4o-Transcribe	0.2176	0.1782	0.1352	0.1372
Cohere Transcribe	0.2582	0.2471	0.1777	0.2088
Backend FT (ours)	**0.1458**	**0.1380**	**0.0911**	**0.1074**

- **BE**: full-audio batch ensemble (Backend FT + Backend Med + reconciler LLM with grounding).
- **MBE**: mini-batch reconciliation (Backend FT + Backend Med, VAD-aligned chunks, 250-token text trailing window, non-revisable).
- **RWE**: rolling-window reconciliation (Backend FT + Backend Med, K=3 buffer window, revisable trailing context).

S, S+PP, B+PP share Backend FT as the sole ASR; their comparison isolates what the pipeline buys on top of streaming and what batch buys over streaming. BE, MBE, RWE add Backend Med and the multi-source reconciler.

5 Results

5.1 Domain-adapted ASR vs. off-the-shelf baselines

Prior to evaluating our full pipeline, we benchmark Backend FT's raw output against two off-the-shelf ASR baselines on the same 226-encounter eval: GPT-4o-Transcribe [5] (closed-source commercial) and Cohere Transcribe [6] (open-weights release). Cohere Transcribe was the top-ranked system on the Hugging Face Open ASR Leaderboard [22] at the time of evaluation. GPT-4o-Transcribe is OpenAI's leading speech-to-text offering. Backend FT is the domain-fine-tuned encoder-decoder ASR introduced in §3.1. The specific model weights and fine-tuning data are proprietary and not disclosed, but do not affect the architectural claim evaluated here.

Backend FT wins on every metric: WER is −33% relative vs. the closed baseline and −44% vs. the open one, with NormWER and KER tracking. Domain adaptation is the substrate the rest of the pipeline builds on. We proceed with Backend FT as the primary backend in all configurations below.

5.2 Post-processing, multi-source reconciliation, revisable context

Table 3 reports the full pipeline progression. Each step in the design space results in a measurable improvement, and the multi-source reconciliation configurations (BE, MBE, RWE) outperform the single-backend ones by a substantial margin.

Three observations are significant:

(1) Putting an LLM in the loop on a single ASR provides significant improvement. S → S+PP cuts KER by ∼57% relative and Norm. KER by ∼62% relative. WER also drops sharply by ∼45% relative. Part of this delta is due to the reconciler's ability to settle streaming-protocol artifacts e.g. spoken-punctuation tokens ("comma", "period").

(2) The streaming/batch gap is small once an LLM is in the loop. S+PP is within 0.5 pp of B+PP on KER. The reconciler closes

Table 3: Metrics for the 226-encounter eval, lower is better.

Config	WER	KER	Norm. WER	Norm. KER
S	0.2760	0.2195	0.1568	0.1547
S+PP	0.1520	0.0951	0.0895	0.0594
B+PP	0.1521	0.0904	0.1003	0.0602
BE	**0.1135**	0.0634	0.0723	0.0396
MBE	0.1462	0.0721	0.0912	0.0431
RWE	0.1136	**0.0603**	**0.0677**	**0.0360**

Table 4: Comparison between MBE, W-MBE (revisable context + more audio context), and RWE. RWE clearly wins.

Metric	MBE	W-MBE	RWE	Rel.	95% CI on RWE−MBE
WER	0.1462	0.1466	0.1136	−22.3%	$[-0.038, -0.027]^{***}$
KER	0.0721	0.0656	0.0603	−16.3%	$[-0.016, -0.007]^{***}$
Norm. WER	0.0912	0.0916	0.0677	−25.8%	$[-0.029, -0.018]^{***}$
Norm. KER	0.0431	0.0414	0.0360	−16.6%	$[-0.011, -0.003]^{***}$

most of the gap that a raw streaming/batch comparison would otherwise show.

(3) Multi-source reconciliation buys more than single-source post-processing. BE drops KER from 0.0904 (B+PP) to 0.0634 (−30% relative; 95% bootstrap CI on the absolute drop $[-0.033, -0.022]$) and Norm. KER from 0.0602 to 0.0396 (−34% relative; CI $[-0.026, -0.016]$). All three multi-source configurations (BE, MBE, RWE) outperform B+PP on every reported metric, with 95% bootstrap CIs excluding zero in every case ($p < 10^{-3}$).

5.3 Within-family ablation: revisable vs. non-revisable trailing context

Table 4 drills into the RWE-vs-MBE comparison as a within-family architectural ablation. Backends, reconciler, and structured grounding are held constant. The architectural difference between MBE and RWE bundles two properties—a wider audio dispatch ($K = 3$ buffers per tick rather than a single VAD chunk) and a liberal prompt allowing the revision of previously-processed text using more acoustic evidence. We add a third configuration - **W-MBE** (wide-window mini-batch): RWE's $K = 3$ audio dispatch with the LLM instructed to reproduce previously-processed text verbatim, matching MBE's PP-edit license on output.

RWE outperforms MBE by 16–26% relative on every metric. The W-MBE column decomposes this win into its two underlying mechanisms. Widening the audio window alone (MBE → W-MBE) modestly improves KER: the $K = 3$ dispatch gives the reconciler more acoustic evidence to disambiguate clinical vocabulary, even when the LLM is conservative with re-writing previous output. The full RWE win is unlocked only when prompt-licensed revision is layered on top of the wider dispatch. ASR errors cluster at *chunk boundaries*, where the streaming pipeline must commit before the disambiguating audio arrives. The wider window supplies that evidence and the revision lets the LLM replace the boundary error.

Qualitative illustration. A clinician dictates: *"The patient has Crohn's disease, treated with infliximab."*

(1) **Streaming commits a phonetic mishearing.** Backend FT, on the short opening chunk, falls back to the homophone "crones" for the unfamiliar "Crohn's." **Streaming output:** *"the patient has crones disease."* No mechanism to revise this commit.
(2) **Rolling-window re-passes with broader audio.** The K=3 buffer extends the audio span to include "…treated with infliximab" - a drug for inflammatory bowel disease. With the medical-dictation language model now primed by the longer prosodic frame, Backend FT re-segments and Backend Med agrees in parallel. **Backends' tick-3 output:** *"the patient has Crohn's disease, treated with infliximab."*
(3) **Reconciler arbitrates.** With both backends in agreement on a recognized clinical term, the reconciler outputs: *"The patient has Crohn's disease, treated with infliximab."*

The disambiguating signal—the surrounding clinical-dictation context that primes the ASR's medical vocabulary—lived in audio bytes the streaming pipeline never re-examined. See Appendix A for three failure modes with this structure.

Latency cost. The quality win comes at a measurable per-tick cost. RWE's median per-tick wall-clock is 2456 ms vs. 1783 ms for MBE—a ∼0.7 s/tick premium (∼38% slower at the median); p_{95} is essentially tied (4657 ms vs. 4779 ms). The reconciler call dominates the gap, driven by more generated tokens. Thus, per-tick latency optimizations should target the reconciler, not the STT models.

5.4 Operating-mode tradeoffs

The four modes occupy distinct points on a quality/latency Pareto frontier (Table 1):

- **Full-batch multi-source (BE)** delivers the highest accuracy but no incrementality. Appropriate for asynchronous workflows (note review, transcription review).
- **Rolling-window (RWE)** matches batch quality on normalized metrics while remaining incremental, at ∼2.5 s/tick wall-clock. Its distinguishing property is revisable context.
- **Mini-batch (MBE)** is the lower-latency incremental option (∼1.8 s/tick wall-clock) at the cost of frozen trailing context.
- **Streaming + reconciliation (S+PP)** delivers low first-emit latency and comes within ∼0.5 pp of B+PP on KER.

A practical pattern emerges: pair pure streaming for in-session UX with a **"final punch"** rolling-window or batch pass at end-of-utterance. Streaming gives the clinician immediate visual feedback, while the final pass finalizes the transcript.

6 Ablations

6.1 Prompt-Transfer: It's the Architecture, Not the Prompt

A possible explanation for the rolling-window win is that the rolling-window prompt is better-engineered than the mini-batch prompt.

Experiment. We graft the rolling-window prompt's single-occurrence / mid-utterance-restart consolidation rule into the mini-batch system prompt and match its hard `max_tokens` cap (derived per call from input length, a structural defense against bulk re-rendering). The purpose of this exercise is to create a

clean test of the prompt-level features that produce coherent multi-chunk output. We ran a 3-trial 226-encounter eval and compared against a fresh 3-trial mini-batch baseline on the same eval set and reconciler-LLM snapshot.

Table 5: Prompt-transfer ablation: RWE prompt features ported into the non-revisable MBE architecture. Δ is computed against the MBE row of Table 3. The gap from MBE to RWE (Table 4) is not closed by porting the prompt.

Metric	PT-MBE	Δ vs MBE	Rel.	95% CI on Δ
WER	0.1421	−0.0041	−2.8%	$[-0.006, -0.002]^{***}$
KER	0.0715	−0.0006	−0.8%	$[-0.003, +0.001]^{ns}$
Norm. WER	0.0864	−0.0048	−5.3%	$[-0.007, -0.003]^{***}$
Norm. KER	0.0406	−0.0025	−5.9%	$[-0.005, -0.001]^{***}$

Result. The prompt-transferred mini-batch shows only small deltas vs MBE: raw KER is statistically indistinguishable, and the other three metrics show 3–6% relative improvements (absolute deltas 0.003–0.005) that are statistically significant under 3-trial bootstrap but small in absolute terms. **Porting the prompt does not move the mini-batch architecture toward the rolling-window operating point.** For reference, the rolling-window architecture (Table 3, RWE row) outperforms MBE—and by extension PT-MBE—by 16–26% relative on every metric (Table 4)—an effect 3–8× larger than the prompt-transfer delta.

This further re-inforces that the rolling-window architecture's quality win comes from the architectural property of revisable trailing context, not the prompt language. This is consistent with the structural argument in §3.4 that chunk-boundary errors in mini-batch cannot be recovered no matter how the per-chunk reconciliation is prompted, because the relevant audio is not in scope.

6.2 Structured Grounding Sensitivity

Removing the structured patient/encounter context block from the reconciler prompt while holding everything else constant regresses the names category by ∼ 5× (from ∼11% to ∼50% keyword error rate on names). Other entity categories are less affected. Even though the block is a few hundred tokens it is the system's only source of information about who the patient and their clinical history. It is the grounding signal that lets the LLM choose the better of the transcript hypotheses when both backends mistranscribe. For LLM systems operating on noisy sensors, this points to a general rule - *the LLM needs a grounding source neither input modality provides.*

6.3 Backend Composition

We swept the multi-source configurations over 2-backend and 3-backend compositions, holding the architecture and reconciler constant. Two findings:

(1) **The medical-specialized backend is load-bearing.** Removing Backend Med from a 3-backend config and replacing with 2 general-purpose backends regresses Norm. KER below the 2-backend baseline (4.54% vs. 4.48%).
(2) **Adding a third backend buys ∼11% relative improvement.** The best 3-backend config reaches Norm. KER 3.99% vs. 4.48% for the 2-backend baseline, well above noise.

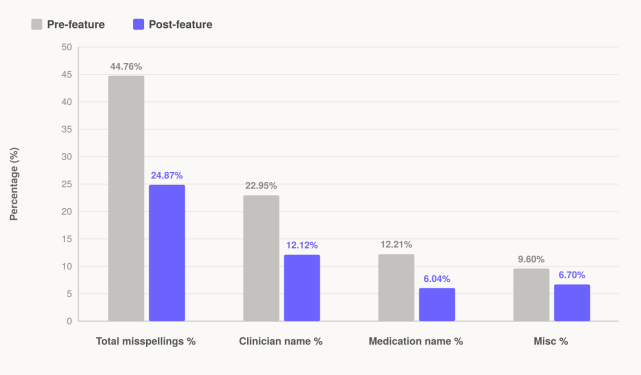

Figure 2: Clinician edit telemetry, pre-feature vs. post-feature correction rates across error categories.

The numbers in this sweep come from a separate 3-backend study with a different ASR/VAD snapshot; rows within the sweep are internally comparable, but absolute values differ from Table 3.

7 Production Validation

We monitor clinician edit telemetry on deployed notes. Over a pre-feature window (five weeks, prior to rollout of the LLM-reconciliation pipeline) versus a post-feature window (fourteen weeks following rollout, spanning 20 medical specialties and 218 clinicians), the rate of misspelling-category correction edits dropped sharply across categories:

- **Medication-name correction rate:** −51% **relative** (12.21% → 6.04%; 192/1573 pre vs. 202/3342 post; $z = -7.42$, $p < 10^{-12}$). This is the category most directly downstream of the medical-specialized backend's contribution.
- **Clinician-name correction rate:** −47% **relative** (22.95% → 12.12%; 361/1573 pre vs. 405/3342 post; $z = -9.77$, $p < 10^{-22}$). Probes the structured-grounding contribution—the structured patient/encounter context block contains clinician identity.
- **Aggregate misspelling correction rate:** −44% **relative** (44.76% → 24.87%; 704/1573 pre vs. 831/3342 post; $z = -14.04$, $p < 10^{-44}$).

The category breakdown is informative (Fig. 2): medication and clinician-name corrections—the categories the architecture is specifically designed to improve—dropped fastest, while generic miscellaneous edits (9.60% → 6.70%) dropped least.

Harmful-correction audit. The structured-grounding block could in principle bias the reconciler towards wrongly "correcting" a term. To bound this empirically, we audit 528 deployed corrections across 87 sampled patient-provider encounters against ground-truth labels. Precision was 99.1% (523/528). The false positives reflect the hypothesized failure mode: the reconciler chose a phonetically plausible medication that did not match the clinician's intent. Representative cases include: *nartriptyline* → *nortriptyline*, *Oculizumab* → *Ocrevus*, *colobizem* → *clobazem*, and *metoprololone* → *metoprolol*.

Caveats. The pre-feature window is bounded by when the LLM-reconciliation pipeline shipped. The post-feature window

accumulated data through fourteen weeks of deployment. We use the fourteen-week post-feature window as the headline because the larger denominator yields tighter point estimates and demonstrates durability beyond launch effects. As a robustness check, we re-compute the post-feature window at a symmetric five-week mark: medication-name 4.93%, clinician-name 15.08%, aggregate misspelling 26.76%, miscellaneous 6.74%. These agree within 1–3 percentage points with the fourteen-week numbers, ruling out user-mix drift as their explanation. The metric is correction edits per note, not total time saved. Mapping editing burden onto clinician efficiency requires additional measurement we have not done. We therefore frame this as evidence of *deployed accuracy improvement on clinically meaningful error classes that the architecture specifically targets*, not as a direct efficiency claim.

8 Implications for LLM-Driven Healthcare Systems

Revisable context as a first-class design variable. The prompt-transfer ablation surfaces a result we believe generalizes: for LLM-in-the-loop systems, the *scope of revisable context* is a first-class design variable, distinct from prompt design and largely orthogonal to model choice. Many production LLM systems today are append-only by default, where each call produces output that is committed downstream. There exist patterns of self-correction methods [23], where an LLM revises its own output through iterative critique, but they do not use external input. By constrast, our results suggest substantial latent quality is recoverable by making prior outputs revisable when fresh evidence is available.

Multi-source LLM reconciliation as an architectural choice. Our architectural contribution consists of LLM reconciliation from heterogeneous ASR backends, grounded in structured patient context. Unlike other ensemble approaches, we frame the system as *multi-source reconciliation grounded in external context.* The LLM integrates parallel noisy signals, using a grounding prior (medications, problems, identity) that neither input source provides.

Real-time constraints shape the upstream design choice. For an LLM-driven note generator that depends on the upstream transcript, the upstream system's chunk size effectively governs how much context any downstream reasoning step gets per call. Treating this as a single end-to-end design problem rather than as "ASR latency" separate from "LLM quality" surfaces tradeoffs that get lost when each subsystem is benchmarked in isolation.

9 Limitations and Future Work

Reproducibility and IP disclosure. Backend identities are partially disclosed for IP reasons. Architecture families can be found in §3.1, but the specific pretrained bases, fine-tuning data, and checkpoints are proprietary. This limits checkpoint-level replication but permits independent reproduction of the architectural design and the multi-source-reconciliation pattern with substitute backends of the same families. The off-the-shelf baselines compared in §5.1 are named commercial products and remain replicable.

Clinical safety and human review. The deployment described in §7 operates as a annotator-in-the-loop system: every transcript produced by the pipeline is reviewed by a trained transcriptionist before clinical use. The architecture and metrics we report describe the signal quality this reviewer operates on. The accuracy improvements documented in §7 therefore reduce reviewer correction burden rather than removing human oversight. A formal clinical safety analysis is a complementary direction left as future work.

Evaluation scope. Our KER computation relies on LLM-based entity extraction; §4 discusses extractor validity (taxonomy designed with medical experts, same-extractor cancellation, WER cross-check, and a direct 122-encounter audit against human labels with pooled F1 0.865 under boundary-relaxed matching—within the published F1 range for clinical NER systems [24]). The production validation in §7 is a deployed-accuracy claim, not a direct efficiency claim; mapping correction-edit reductions onto clinician time savings requires additional instrumentation we have not done.

External validity and generalization. The quantitative claims rest on a constrained evidence base. The 226-encounter eval is an in-house clinical-dictation corpus rather than a sample from a deployment stream. The production validation in §7 comes from a single health-system deployment. Cross-institution transfer to other clinician populations and acoustic capture conditions (e.g., various microphones) is unmeasured. The eval is English-only with a non-comprehensive accent distribution, and lacks within-specialty performance variance analysis despite spanning 20+ specialties.

Future work. Four directions extend the present results: (1) broadening the evaluation set across more speakers, accents, and short-utterance dictations; (2) ensembling additional model combinations beyond the Backend FT + Backend Med configuration deployed here; (3) fine-tuning the LLM reconciler on the in-domain reconciliation task to reduce per-tick latency for the rolling-window architecture; (4) a public demonstration of the architecture on a shareable clinical corpus to enable replication and community benchmarking.

10 Conclusion

We present a production clinical-dictation pipeline that treats the transcript as revisable evidence rather than append-only text. The system runs heterogeneous ASR backends in fixed parallel dispatch and passes their raw hypotheses, together with structured patient context, to an LLM reconciler. A rolling K-buffer audio re-pass gives the reconciler revisable trailing context, allowing earlier transcript decisions to be corrected when later audio resolves them. On a 226-encounter clinical dictation evaluation, multi-source reconciliation is the dominant gain, reducing normalized keyword error by 40% relative to a single fine-tuned backend with post-processing. The revisable rolling-window variant adds a further raw-quality gain over the non-revisable mini-batch ensemble, reducing raw WER by 22.3% and raw KER by 16.3%. In deployment, medication-name correction edits dropped by 51%. The central design lesson is that revisability is not a prompting trick: for LLM-in-the-loop healthcare systems built on noisy upstream signals, the scope of prior output that the model may revisit is a first-class architectural choice.

References

[1] C. Sinsky, L. Colligan, L. Li, M. Prgomet, S. Reynolds, L. Goeders, J. Westbrook, M. Tutty, and G. Blike. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760, 2016.

[2] L. R. Willoughby, D. J. Hekman, and B. H. Schnapp. The effect of dictation on emergency medicine resident time to note completion. *Western Journal of Emergency Medicine*, 26(6):1499, 2025.

[3] Polaris Market Research. Voice technology in healthcare market size, share & trends analysis report, 2024–2032, 2024.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[5] OpenAI. Introducing Next-Generation Audio Models in the API. OpenAI product release, 2025. Accessed 2026-05-19.

[6] J. Mack, E. Ranjan, W. Beller-Morales, B. Venkitesh, and P. Richemond. cohere-transcribe-03-2026 (Revision d96e814). Hugging Face, 2026. DOI: 10.57967/hf/8653.

[7] Health AI Developer Foundations. A conformer-based medical speech-to-text model for clinical dictation. 2025.

[8] Nuance Communications (Microsoft). Dragon medical one: Clinical speech recognition. Product documentation, 2024.

[9] Speechmatics. The next generation of AI medical speech recognition is here. Company blog post, 2025.

[10] T. Afonja, T. Olatunji, S. Ogun, N. A. Etori, A. Owodunni, and M. Yekini. Performant ASR Models for Medical Entities in Accented Speech. In *Proc. Interspeech 2024*, pages 2315–2319, 2024.

[11] J. Li. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022. Now Publishers.

[12] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of IEEE ASRU Workshop*, pages 347–354, 1997.

[13] C. Chen, Y. Hu, C.-H. H. Yang, S. M. Siniscalchi, P.-Y. Chen, and E.-S. Chng. HyPoradise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36:31665–31688, 2023.

[14] R. O'Kane, D. Stonehouse-Smith, L. C. U. Ota, R. Patel, N. Johnson, C. Slipper, J. Seehra, S. N. Papageorgiou, and M. T. Cobourne. Transcription accuracy of automatic speech recognition for orthodontic clinical records. *Journal of Dental Research*, 2025.

[15] D. Macháček, R. Dabre, and O. Bojar. Turning Whisper into real-time transcription system. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 17–24, 2023.

[16] M. Bain, J. Huh, T. Han, and A. Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *INTERSPEECH*. ISCA, 2023.

[17] J. Prakash, B. Kumar, K. Hacioglu, B. Sharma, S. Gopalan, M. Chetlur, S. Venkatesan, and A. Stolcke. Better Pseudo-labeling with Multi-ASR Fusion and Error Correction by SpeechLLM. In *Proc. Interspeech 2025*, pages 579–583, 2025.

[18] H. Wang, G. Hu, G. Lin, W.-Q. Zhang, and J. Li. Simul-Whisper: Attention-Guided Streaming Whisper with Truncation Detection. In *Proc. Interspeech 2024*, pages 4483–4487, 2024.

[19] K. Binici, A. R. Kashyap, V. Schlegel, A. T. Liu, V. P. Dwivedi, T.-T. Nguyen, X. Gao, N. F. Chen, and S. Winkler. MEDSAGE: Enhancing robustness of medical dialogue summarization to ASR errors with LLM-generated synthetic dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23496–23504, 2025.

[20] C.-H. H. Yang, T. Park, Y. Gong, Y. Li, Z. Chen, et al. Large language model based generative error correction: A challenge and baselines for speech recognition, speaker tagging, and emotion recognition. In *IEEE SLT*, 2024.

[21] Y.-T. Lin, Z. Chen, P. Zelasko, Z. Wan, X. Yang, Z.-C. Chen, K. C. Puvvada, S.-W. Fu, K. Hu, J.-W. Chiu, J. Balam, B. Ginsburg, Y.-C. F. Wang, and C.-H. Yang. NeKo: Cross-modality post-recognition error correction with tasks-guided mixture-of-experts language model. In *ACL Industry Track*, 2025.

[22] Hugging Face. Open ASR Leaderboard. https://huggingface.co/spaces/hf-audio/open_asr_leaderboard, 2026.

[23] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

[24] M. C. Durango, E. A. Torres-Silva, and A. Orozco-Duque. Named entity recognition in electronic health records: A methodological review. *Healthcare Informatics Research*, 29(4):286–300, 2023.

A Worked Examples from the Eval Set

The three examples below are drawn from the 226-encounter eval and quote actual reference, S+PP, and RWE outputs from those encounters. They share a common property: the streaming pipeline's single-pass single-backend ASR produces low-information outputs at chunk boundaries, in low-SNR audio, and when faced with proper nouns or domain terms—and the downstream LLM post-processor cannot reliably correct what it cannot distinguish from valid input. RWE addresses each: (1) two parallel ASR backends, one medical-specialized, vote on each utterance; (2) each speech segment is re-transcribed across K overlapping windows so context for one segment is available when reconciling the next; (3) the reconciliation prompt explicitly treats the second-backend output and the prior-tick output as cross-references, letting the LLM reject content that fails to align across passes.

A.1 Background-speech intrusion (encounter `0b867859`)

Reference (gold).

> *Samaniego Anderson is a 2-month-old male with no past medical history who presents with increased work of breathing and congestion for the past 2 days. Patient was recently exposed to some sick contacts in her home. Patient is also, up until this point, healthy. Birth was normal with no perinatal complications. Patient did not require any stay in the neonatal ICU. For the past 2 days, patient has had notable increased work of breathing. Mom has tried using bulb suction and Frida nasal deep suction with minimal success. Patient has had also some relief with the use of a humidifier in the home. Mom denies any fevers, chills, cyanosis, change in behavior.*

S+PP (WER = 0.313).

> *[Inaudible] Anderson is a 2-month-old male with no past medical history who presents with increased work of breathing and congestion for the past 2 days. Patient was recently exposed to some sick contacts in her home. Patient is also up until this point healthy. Birth was normal with no perinatal complications. Patient did not require any stay in the neonatal ICU. Did you see my speech? For the past 2 days, patient has had notable increased work of breathing. Mom has tried using bulb suction and Frida nasal deep suction with minimal success. Patient has had. Thank you. Because I've never been so scared in my entire life. Also some relief with the use of a humidifier in the home. Mom denies any fevers, chills, cyanosis. Change in behavior. I love you. Okay, you're gonna say something right now. Mm.*

RWE (WER = 0.000).

> *Samaniego Anderson is a 2-month-old male with no past medical history who presents with increased work of breathing and congestion for the past 2 days. Patient was recently exposed to some sick contacts in her home. Patient is also, up until this point, healthy. Birth was normal with no perinatal complications. Patient did not require any stay in the neonatal ICU. For the past 2 days, patient has had notable increased work of breathing. Mom has tried using bulb suction and Frida nasal deep suction with minimal success. Patient has had also some relief with the use of a humidifier in the home. Mom denies any fevers, chills, cyanosis, change in behavior.*

Failure modes. S+PP decoded background and cross-talk audio (TV, a child, someone speaking elsewhere in the exam room) as part of the dictation stream—"Did you see my speech?", "Thank you. Because I've never been so scared in my entire life.", "I love you. Okay, you're gonna say something right now." The post-processor cannot reliably filter these inserted phrases because they are grammatical English emitted by the ASR with no marker distinguishing them from the clinician's speech. The patient's last name also drops to `[Inaudible]`. RWE routes the same audio through Backend FT and Backend Med and re-transcribes each speech segment across K overlapping windows; the reconciler sees Backend Med carrying the clinical narrative while the cross-talk phrases appear in only one backend's output and are dropped.

A.2 Drug-name garble, lost negation, sentence fragmentation (encounter `baf0ba0f`)

Reference (gold).

> *Rogers Conner is a 58-year-old male with history of alcohol use disorder, cirrhosis (MELD-Na 18) with esophageal varices status post banding (6/2024), portal hypertension, who presented with hematemesis and melena. … He denies any prior episodes of melena. No recent NSAID use. Not on any anticoagulation. He otherwise denies any lightheadedness, nausea/vomiting/hematemesis, chest pain, shortness of breath, abdominal pain, dysuria/hematuria, bright red blood per rectum.*

S+PP (WER = 0.202).

> *Rogers Conner is a 58-year-old male with history of alcohol use disorder, cirrhosis (MELD-Na 18), with esophageal varices. Test post pending (6/2024), portal hypertension, who presented with hematemesis and melena. … He denies any prior episodes. Of [inaudible]. No recent. And said use. Not on any anticoagulation. He otherwise denies any lightheadedness. Come on, nausea, vomiting, hematemesis, chest pain, shortness of breath. Calm abdominal pain, dysuria/hematuria, bright red blood, pruritus..*

RWE (WER = 0.025).

> *Rogers Conner is a 58-year-old male with history of alcohol use disorder, cirrhosis (MELD-Na 18), with esophageal varices status post banding (6/2024), portal hypertension, who presented with hematemesis and melena. … He denies any prior episodes of melena. No recent NSAID use. Not on any anticoagulation. He otherwise denies any lightheadedness, nausea/vomiting/hematemesis, chest pain, shortness of breath, abdominal pain, dysuria/hematuria, bright red blood per rectum.*

Failure modes. Four distinct clinical-safety issues compound in one note. (1) "status post banding"—a specific surgical procedure—is decoded as "Test post pending", a phrase that not only is grammatically wrong but suggests an outstanding test rather than a completed procedure. (2) "NSAID" is decoded as "And said"—the drug-class term that is part of the GI-bleed differential is lost entirely. (3) The sentence "No recent NSAID use"—which rules out NSAIDs as a cause of the bleeding—is fragmented across three half-sentences ("No recent. And said use.") with the negation orphaned from its object. (4) Background utterances "Come on" and "Calm" are inserted inline, and "pruritus" is hallucinated into the negative-review list (the patient is not denying pruritus; the word does not appear in the reference at all). RWE recovers all four: Backend Med correctly emits "NSAID" and "status post banding," and the reconciler rejects the non-clinical interjections by content alignment with the surrounding clinical context.

A.3 Wrong drug name and cross-document hallucination (encounter `7514a12d`)

Reference (gold). Assessment-and-plan section:

> *… Assessment:*
> *1. PCI to the mid-LAD.*
> *Plan: Currently on statin as well as Plavix and Brilinta.*
> *2. Bilateral CVA.*
> *Plan: Continue Brilinta and Plavix. Patient admit to telemetry to evaluate for any underlying atrial fibrillation.*
> *3. Syncopal episode. Syncopal episode likely in the setting of vasovagal, likely not in the setting of stroke.*
> *4. Hyperlipidemia. Continue high-intensity statin.*

S+PP (WER = 0.670).

> *… Assessment. PCI to the mid LAD. Americans to, I quote, indulge in modern. Plan currently wants to end as well as Plavix and Burlington. Second problem. Bilateral adaptive, which is CVA. Plan, continue. Brilinta and Plavix? Patient admitted to telemetry to evaluate for any underlying atrial fibrillation. 3, syncopal episode. Sync [inaudible] likely not in the setting of stroke? For consumers or people that have a hyperlipidemia, continue high-intensity statin. That they were eating. So on January 7, 2026, the food pyramid was at-*

RWE (WER = 0.187).

> *… Assessment. PCI to the mid LAD. Plan. Currently on statin as well as Plavix and Brilinta. Second problem, bilateral CVA. Plan. Continue Brilinta and Plavix. Patient admitted to telemetry to evaluate for any underlying atrial fibrillation. 3, syncopal episode. Syncopal episode likely in the setting of vasovagal, likely not in the setting of stroke. 4, hyperlipidemia. Continue high-intensity statin.*

Failure modes. "Brilinta"—a specific antiplatelet agent—is decoded as "Burlington" on first mention (a city name; a clinically dangerous medication substitution if entered into the chart). The note is interleaved with apparent stray-audio decodings unrelated to clinical content: "Americans to, I quote, indulge in modern.", "That they were eating. So on January 7, 2026, the food pyramid was at-"—content that would have to be detected and stripped by downstream human review. Several clinical phrases are fragmented or restructured into questions ("Brilinta and Plavix?", "likely not in the setting of stroke?"). RWE's Backend Med carries "Brilinta"

correctly on both mentions, and the reconciler filters the stray non-clinical decodings by content alignment with the surrounding medical narrative.